Unsupervised Landmark Discovery via Self-Training Correspondence

Dimitrios Mallis

Computer Vision Laboratory

School of Computer Science

University of Nottingham

This dissertation is submitted in partial fulfillment of the conditions for the award of the degree of *Doctor of Philosophy.*

February 2023

Abstract

Object parts, also known as landmarks, convey information about an object's shape and spatial configuration in 3D space, especially for deformable objects. The goal of landmark detection is to have a model that, for a particular object instance, can estimate the locations of its parts. Research in this field is mainly driven by supervised approaches, where a sufficient amount of human-annotated data is available. As annotating landmarks for all objects is impractical, this thesis focuses on learning landmark detectors without supervision. Despite good performance on limited scenarios (objects showcasing minor rigid deformation), *unsupervised landmark discovery* mostly remains an open problem. Existing work fails to capture *semantic landmarks*, i.e. points similar to the ones assigned by human annotators and may not generalise well to highly articulated objects like the human body, complicated backgrounds or large viewpoint variations.

In this thesis, we propose a novel *self-training* framework for the discovery of unsupervised landmarks. Contrary to existing methods that build on auxiliary tasks such as image generation or equivariance, we depart from generic keypoints and train a landmark detector and descriptor to improve itself, tuning the keypoints into distinctive landmarks. 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. Our detector can discover highly semantic landmarks, that are more flexible in terms of capturing large viewpoint changes and out-of-plane rotations (3D rotations). New state-of-the-art performance is achieved in multiple challenging datasets.

Acknowledgements

This PhD and the incredible personal journey that came along could never be possible without my supervisor Dr Yorgos Tzimiropoulos. For all the time and effort he spent guiding me, for the trust he showed in me, for pushing me when I needed it, for showing me how it's done, for all this and much more, Yorgos, I sincerely thank you.

Besides, special thanks to my second advisor Dr Matt Bell for his persistent guidance and constant support. I would also like to thank my coauthor Dr Enrique Sanchez whose input, feedback, and suggestions were instrumental in developing this work.

Thanks to all the fantastic people of the Computer Vision Lab. Keerthy, Ioanna, Aaron, Jing, Siyang, Mani, Johann, Zane, Joy and Bowen for their encouragement, support, inspired conversations and late evening jokes at B46. A special mention to Aaron for all the technical problems he simply made disappear. Thanks to all the CVL professors, Tony Pridmore, Michel Valstar, Andrew French and Mike Pound, for nurturing an excellent research environment. I also want to thank the Douglas Bomford Trust that enabled me to pursue a PhD through its generous funding. Thanks to my professors Dr Nikos Mitianoudis and Dr Avi Arampatzis, for guiding me through my first steps in research.

Thanks to my friends who are always and will always be here for me since childhood. Thanks to my relatives, especially my parents, sister, and grandmother, where no words are needed to express my gratitude. Finally, thanks to Vanessa for taking this amazing journey alongside me.

Dedicated to my grandfather, whose memory will always live in my heart.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

Understanding the pose of arbitrary objects is an integral part of visual perception. Physical objects can be highly deformable, able to articulate into complicated part configurations. Accurate estimation of their structure constitutes a long-standing Computer Vision problem. Pose estimation requires a deeper understanding of an object's intrinsic properties as well as accidental factors like viewpoint or illumination.

Given the significance of the task, various models for representing the pose of physical objects have been proposed. Notable examples are earlier part-based models like Pictorial Structures [51] and Deformable Part Models [49] where an object is represented as a collection of parts, hierarchical bone representations [156] and more recently dense pose representations [60] where each pixel of the object is mapped to a dense template grid. The most widely used pose model though, is the *skeleton-based model*. The skeleton-based model is a simple and flexible representation, where pose is expressed as a collection of K point coordinates tracking particular semantic locations on the object of interest. We will refer to these locations as *object landmarks* and the task of detecting them as *landmark detection (or localisation)*. A visual example of object landmarks detected on facial images

Figure 1.1: Object pose captured as a set of K coordinates or *object landmarks*. Landmarks are detected through our unsupervised landmark detection approach that will be introduced in this thesis.

can be seen in Fig. 1.1.

Landmark detection is a computer vision task on which there is extensive literature. Earlier landmark detectors where based on Active Appearance Models [36] or Cascaded Regression [43]. More recently, deep learning based approaches have achieved impressive performance in challenging, in-the-wild datasets like COCO [105]. Deep landmark detectors are commonly trained in a *supervised* manner on large datasets of manually annotated images. For example, the COCO dataset includes approximately $250,000$ person instances manually labelled for 17 object landmarks. Under this supervised setting, recent deep models like Hourglass [122] or HRNet [155] are able to push the envelope on landmark detection performance significantly.

This rapid progress has led to a wide range of applications for landmark localisation technology, particularly around the *human-sensing* domain, i.e. problems related to the extraction of information regarding people present in an environment. Human Pose Estimation (*HPE*) find several use cases ranging from surveillance to action recognition, autonomous driving, gaming and animation. Alignment of the human face (i.e. facial point localisation) is a fundamental step in most facial analysis systems. Typical applications include security, for example, face recognition for authentication, healthcare (emotion recognition, pain assessment, e.t.c) and, more recently, face animation and reenactment. Estimating the shape of the human hand is essential for gesture recognition and control or sign language understanding.

Compared to the success achieved in human sensing, development of deep landmark

detectors for arbitrary object categories has attracted far less attention. Animal Monitoring, for example, is a domain where such technology could facilitate industrial progress. Animal pose estimation is required for the automation of various tasks, from behaviour recognition to locomotion, lameness or oestrus detection. Pose information has been used in pain assessment [67] and monitoring of the farrowing process for pigs [126]. Regardless of the apparent potential, only a few animal landmark detectors have been recently proposed [79, 191, 101] compared to the vast literature on human pose estimation and face alignment. Overall, vision technologies are rarely used in industrial farms. Instead, information about the animal's state is commonly recovered through various technologies like neck mounted collars and other identification devices.

This thesis identifies that the ability to estimate the pose of arbitrary objects can lead to a broad and diverse range of applications. Further to the aforementioned case of animal monitoring, some innovative use-cases explored in recent literature include arbitrary object animation [150], object control for Reinforcement Learning in [91] and class-dependent video prediction [82]. So why isn't there increasing effort towards the development of deep landmark detectors for more object categories? Training of such models requires large amounts of annotated data. Large scale manual annotations are expensive and time-consuming to collect, constraining most existing work around the human sensing domain where annotated datasets are already available.

Thus, this thesis formulates the following research question. *"Can we develop robust landmark detectors for arbitrary object categories with minimum requirement for expensive manual annotations"*. As we will discuss later in this Chapter (Subsection 1.5), we address this problem statement by solving the following: *(a)* we learn a landmark detector directly from raw images without manual supervision, *(b)* a robust method is developed, able to account for object deformations and out-of-plane rotations. *(c)* we detect landmarks with high semantic value (points that track object locations similar to the ones assigned by human annotators).

1.2 Motivation

The most significant limiting factor for training deep landmark detectors on arbitrary object categories is the requirement for large annotated datasets. Characteristically the COCO dataset required approximately $70k$ work hours [105] to develop. Fine-grained annotations like keypoint annotations or segmentation masks are particularly time-consuming (compared to class labels), with the average annotator requiring several mins per image. The amount of data one needs for learning a deep neural network varies with different tasks. However, even the smaller benchmarks (for human pose estimation or face alignment) commonly include at least $10,000$ annotated images. Such large data collections are currently only available in the human sensing domain (see Fig. 1.2), thus introducing an important barrier for landmark detection on novel object categories (where annotated data are scarce).

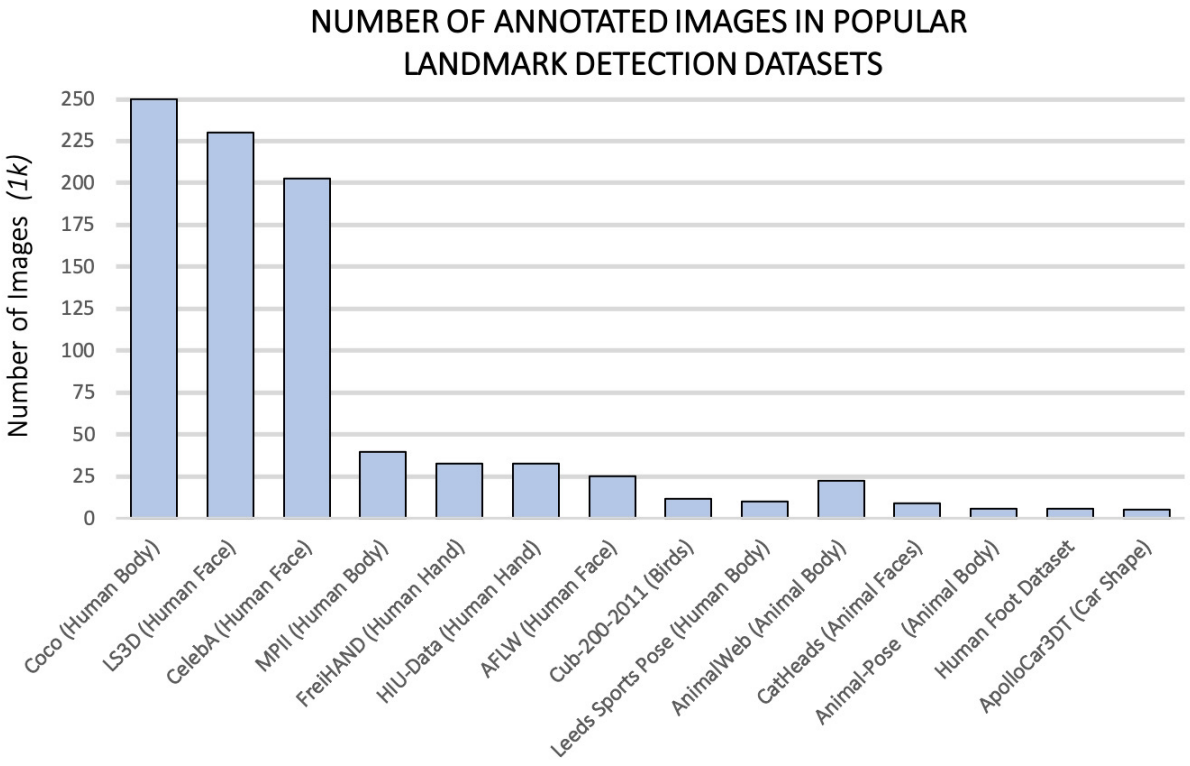


Figure 1.2: Number of manually annotated images in popular landmark detection datasets for various object categories. Notice that larger datasets mostly capture object categories related to *human sensing*, whereas for diverse categories only very few annotated images are publicly available.

The cost of data annotation is not the only limitation of supervised learning. Deep Neural Networks (*DNNs*) trained on labelled datasets have been shown to generalise poorly under domain drift [52, 172] or be sensitive to adversarial examples [161] due to reliance on spurious correlations [171, 107]. These correlations are informative patterns for predicting training samples but do not hold in general. In a supervised setting, certain spurious features correlate with the manual labels, a phenomenon that has also been referred to as dataset bias [35] or group shift [140]. Recent work suggests that this issue can be mitigated in alternative learning paradigms that do not utilise manual supervision [31]. Moreover, human annotators usually recruited through crowdsourcing platforms like Amazon Mechanical Turk (MTurk) are prone to errors or introduction of their own biases to the annotation process [47, 46] raising the need for carefully designed pipelines to ensure annotation quality.

Overall, even though supervised learning has facilitated tremendous advancements in the field of AI in recent years, it also introduces a bottleneck due to the requirement for massive amounts of annotated data. In contrast, human cognition allows learning new skills without immediate supervision for every task. The ability of biological intelligence to learn new concepts through observation and association with previously acquired background knowledge has mostly eluded AI systems. A research direction that attempts to alleviate these limitations is self-supervised learning (*SSL*). In a recent blog post[1], Turing Award winner Yann LeCun referred to SSL as:

> *One of the most promising ways to build background knowledge and approximate a form of common sense in AI systems.*

With self-supervised learning, models can train on orders of magnitude more data, enabling the modelling of more subtle and less frequent patterns. Recently, SSL powered methods have achieved remarkable performance across AI fields. Particularly impressive is the performance of self-supervised Natural Language Models like GTP-3[11] and BERT [41]

[1] `https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/`

whereas progress has also been demonstrated on image classification [56, 54] and speech recognition [146]. The central premise of self-supervised learning is that supervisory signals are obtained from the data itself by learning to solve a particular *pretext task*. Common pretext tasks used in the literature include inferring unobserved or hidden parts of the data, predicting clustering assignments or future frames and learning from augmented views of the same underlying image.

Motivated by the potential of SSL-based methods on various tasks, this thesis will explore self-supervised learning for training a deep landmark detector directly from raw images without manual supervision. We will refer to this task as *unsupervised landmark discovery*[2].

1.3 Main Challenges

From a first glance, unsupervised landmark discovery seems like an impossible task. A human annotator has understanding of the notion of objects and their parts, viewpoint invariance, occlusion and self-occlusion, and examples of which landmarks to annotate at their disposal. Hence, it is completely unclear what pretext task should be chosen for unsupervised landmark discovery with neural networks. Recent methods have focused on two principles/tasks: *equivariance* [164, 163, 162] and *image generation* [200, 72, 144]. Equivariance based methods train a landmark detector to be consistent under known synthetic transformations (on augmented views of the same image). Methods based on image generation or reconstruction commonly condition a generation task on the underlying shape of a source object (expressed as object landmarks).

Even though good performance has been shown, especially for less deformable objects (human and animal faces, shoes, e.t.c), recent methods still suffer from two major limita-

[2]Even though *self-supervised* landmark discovery is a more appropriate term, we will refer to our primary task as *unsupervised* landmark discovery for consistency with recent methods. The term *unsupervised* can be misleading since it suggests the lack of supervision. Self-supervised learning does use supervisory signals that originate from the data itself.

tions. **(1)** Object landmarks are not learned explicitly (through regression or heatmap estimation as it's usually the case with supervised methods). Instead, they are automatically discovered, for example, as an intermediate step of image generation or by invariance to geometric transformations. It is unlikely that by optimising such a proxy objective, one could learn object landmarks with a clear semantic meaning (i.e. landmarks similar to those annotated by humans). **(2)** Landmark detection on raw images is enabled through *synthesised image pairs* since local correspondences for unpaired images are not known in the unsupervised case. Commonly a known transformation is applied to a *single image* in order to create a different view of the same object. The trained network learns to either be consistent under the known transformation [164, 163], or generate the target view from the source view [196, 72, 144]. Learning from synthetic image pairs results in representation with limited robustness to intraclass variation that may not generalise well well to highly articulated objects like the human body, complicated backgrounds or large viewpoint changes (i.e. 3D rotations).

1.4 Contributions

This thesis addresses both these limitations by introducing two novel perspectives for unsupervised landmark discovery. These are *self-training on generic keypoints* and *clustering correspondence*.

Self-training on generic keypoints: Self-training refers to a particular flavor of self-supervised learning where the training model's own inference is used as a training signal. We introduce a novel self-training method for unsupervised landmark detection that uses generic keypoints as initialisation. Generic keypoints do not require manual effort to collect and can enforce a strong prior on the semantic meaning of detected landmarks. Consider that for many object categories, manually annotated landmarks are mostly located on edges/corners of an object's surface. Hence, they could be detected by a generic keypoint detector with good repeatability. In that sense, we can think of

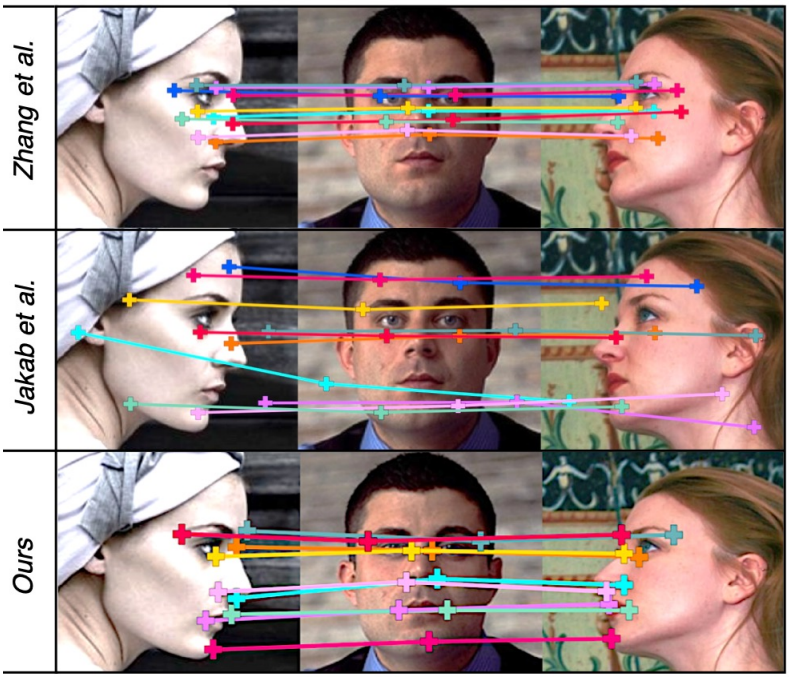


Figure 1.3: Comparison between the landmarks discovered by our approach and those by previous methods. Our approach provides landmarks that can better capture correspondence across large viewpoint changes. For example, there are visible landmarks detected in profile views that are matched with their corresponding points in the frontal view. Moreover, there are detected landmarks that are geometrically consistent in all 3 views. The method of [200] moves the point cloud altogether to fit the face region and hence loses correspondence (e.g. see red or orange point). Also, the method of [72] cannot cope well for such large changes in pose.

the generic keypoints as a noisy mixture containing unsupervised object landmarks along with points in the background and non-corresponding points on the objects of interest. Starting from these generic keypoints, we propose a self-training approach that learns a landmark detector that progressively improves itself by filtering out noisy points and recovering stable object landmarks. Landmark locations are regressed directly through heatmap estimation (not learned as the outcome of a proxy task). Our approach can discover object landmarks similar to those detected from supervised methods for various object categories.

Clustering Correspondence: We propose a novel approach to train an unsupervised landmark detector that is not based on synthetic views of the same object but learns directly from *unpaired images*. Our method is based on deep clustering of local features. We alternate between recovering landmark correspondence through deep clustering and

landmark feature learning without labels using the recovered correspondences. Since correspondences are recovered iteratively through self-training, we can directly use pairs of different images to learn better features. This formulation results in robust learning of landmark representation with stronger invariance to appearance or viewpoint variations. We demonstrate that learning from pairs of different images can enable better localisation performance compared to the use of synthetic transformations.

We combine these novel ideas in an effective framework for unsupervised landmark discovery from raw unlabelled images that do not require any manual supervision and can be applied to arbitrary object categories.

In summary, our main contributions are:

- We are the first to explore *self-training from generic keypoints* for unsupervised landmark discovery (Chapter 4).
- We study for the first time, learning under extreme label noise for fine-grained localisation and propose several techniques to mitigate the effect of noisy annotations (Chapter 4).
- We propose, for the first time, an unsupervised landmark detection method that can learn correspondences from unpaired images via deep clustering. Our approach results in robust landmark representation with stronger invariance to appearance or viewpoint variations (Chapter 5).
- We propose a method that detects highly semantic object landmarks (compared to related work) similar to the ones assigned by human annotators for various object categories (Chapter 4, 5, 6).
- We combine *self-training on generic keypoints* and *clustering correspondence* into a simple and effective framework for unsupervised landmark discovery (Chapter 6).
- We are the first to enable flipping augmentation for the training of unsupervised landmark detection models. This augmentation is not utilised for the unsupervised

case since landmark correspondence after flipping is unknown. We propose a strategy that recovers flipping correspondence via deep clustering (Chapter 6).

- We are the first to develop an unsupervised landmark detector that can capture large changes in 3D viewpoint. Our method provides superior results on a variety of difficult facial and human pose datasets (LS3D [13], BBCPose [24], Human3.6M [69], PennAction [198]), notably without utilizing temporal supervision (Chapter 5, 6).

1.5 Publications

The research presented in this thesis has been partially published in the following conferences and journals, sorted by date:

[114] Mallis Dimitrios, Enrique Sanchez, Matt Bell and Georgios Tzimiropoulos.

Unsupervised Learning of Object Landmarks via Self-Training Correspondence.

Neural Information Processing Systems (NeurIPS), 2020

[115] Mallis Dimitrios, Enrique Sanchez, Matt Bell and Georgios Tzimiropoulos.

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

IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2023.

Chapter 2

Background

This Chapter will first define the landmark detection task. We will outline the standard training pipelines and mention popular network architectures for landmark detection. We will then extend our discussion to the unsupervised case that is the main focus of this thesis and provide a first introduction to the connection between landmarks and generic keypoints. Finally, the standard evaluation frameworks for both supervised and unsupervised landmark detection will be described, along with the various benchmark datasets that will be used for experimentation in later Chapters.

2.1 The Landmark Detection task.

We will define *landmark detection* as the Computer Vision task of localising a set of K coordinates on monocular images of a specific object category (human face, body, hand, e.t.c). This pose representation is often referred to in the literature as the skeleton-based model [203] and the detected points as *object landmarks*. Object landmarks are manually selected during the dataset development stage to capture semantic or highly deformable object locations. For example, body joints (legs, wrists, ankles, e.t.c) for human pose estimation or facial landmarks (corners of the eyes, nose, mouth) for face alignment. More

formally a landmark detector aims to extract a shape vector $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_K^T) \in \mathbb{R}^{2K}$ where $i \in \{1, 2, \ldots, K\}$ and $\boldsymbol{y}_i = (x_i, y_i) \in \mathbb{R}^2$ are the x and y coordinates of the i^{th} object landmark in absolute image coordinates. Without loss of generality, we will assume that the input image $\mathbf{x}$ is normalised w.r.t a bounding box b such as all training samples have approximately the same centre and scale.

2.2 Supervised Landmark Detection.

To estimate the shape vector $\boldsymbol{y}$ one can learn a function $\boldsymbol{\Psi} : \mathcal{X} \rightarrow \mathcal{Y}$ where $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^{W \times H \times 3}\}$ is the image domain. The function $\boldsymbol{\Psi}$ will be a Deep Neural Network (*DNN*) with parameters θ. We will refer to $\boldsymbol{\Psi}$ as the landmark detection network or simply *landmark detector*. Depending on the form of $\mathcal{Y}$ we can identify 2 main variants of pose estimation approaches, that is *regression-based* methods and more recently *heatmap-based* methods. For regression-based methods, $\mathcal{Y}$ has the form $\mathcal{Y} \in \mathbb{R}^{2K}$ and pose estimation is formulated as a end-to-end regression problem where the values of the shape vector are directly estimated. For heatmap-based methods, $\mathcal{Y} \in \mathbb{R}^{W \times H \times K}$ and the goal is to estimate a set of K heatmaps $\{\mathbf{H}_1, \ldots, \mathbf{H}_k\}$ where $\mathbf{H}_i \in \mathbb{R}^{W \times H}$ and the value on $\mathbf{H}_i(x, y)$ indicates the probability that the i^{th} landmark lies in the coordinates (x, y). The shape vector $\boldsymbol{y}$ can get directly infered from the heatmap-based representation simply as the heatmap coordinate with the highest probability or more formally $\boldsymbol{y}_i^* = \underset{(x_i, y_i)}{\operatorname{argmax}} \mathbf{H}_i$.

The function $\boldsymbol{\Psi}$ can be learned from a training set of manually annotated images. A common learning objective is to minimise the Mean-Squared-Error (*MSE*) between predicted and ground-truth shape vectors. For regression-based methods, the MSE loss takes the form:

$$L_{MSE} = ||\mathbf{y} - \boldsymbol{\Psi}(\mathbf{x}, \theta)||^{\mathbf{2}} \tag{2.1}$$

In heatmap-based approaches,

$$L_{MSE} = \sum_{i=1}^{K} \sum_{x,y} ||\mathbf{H}'_i - \mathbf{\Psi}_i(\mathbf{x}, \theta)||^{\mathbf{2}} \tag{2.2}$$

where $\mathbf{H}'_i$ is the groundtruth heatmap formed by placing a small 2D gaussian of constant variance on the (x_i, y_i) groundtruth point.

Even though there is extended literature on regression-based methods [170, 130, 21, 156, 113, 124, 195], these approaches have shown to be less accurate. They suffer from various shortcoming [168] like the unnecessary learning complexity of mapping the RGB image input to an XY coordinate space or the fact that they cannot be naturally extended to allow for multiple occurrences per object landmark (required in multiperson human pose estimation). Heatmap-based landmark detection has recently attracted more attention [134, 168, 186, 122, 155, 32] due to its superior performance. It will be the approach we will follow in this thesis for the unsupervised detection of object landmarks.

2.3 Model Architectures for Landmark Detection.

Convolutional Neural Networks (*CNNs*) have been shown to produce state-of-the-art results for most Computer Vision tasks. Given their impressive performance, they constitute a natural building block for landmark detection networks. Many popular CNN architectures have been explored, commonly for human pose estimation or face alignment. This section will briefly introduce the reader to some popular CNN architectures for landmark detection.

A good starting point is the AlexNet [89] architecture, introduced in 2012 and considered a seminal work in Computer Vision. AlexNet has 62.4 million parameters in total and consists of 5 convolutional layers followed by max-pooling and three fully connected layers at the end of the network. Some of the novelties of this work also include the use of the *ReLU* function instead of *Tanh*, a normalisation layer, as well as overlapping

pooling and dropout. Variations of this architecture were used for landmark localisation [170, 28, 169], most notably the cascaded regression network deemed DeepPose in [170].

With deep networks receiving increased attention following the introduction of AlexNet, even deeper architectures quickly emerged. VGG [151] pushed network depth to 19 layers by reducing the size of the convolutional filters to (3×3) and the stride to (1×1). This reduction led to fewer parameters per layer and allowed for the efficient training of deeper models. VGG is used as the backbone of the popular OpenPose [17] detector, as well as [72, 92, 50] for both human pose estimation and face alignment. GoogleNet [160] was another breakthrough architecture. It comprise a 22 layer deep neural network that is wider in the sense that each layer has multiple parallel convolutions. Its basic block is the inception block that comprises multiple filters of different sizes (1×1, 3×3, 5×5) that operate on the same resolution, followed by a 1×1 filter for dimensionality reduction. A GoogleNet based landmark detector was proposed in [21] for human pose estimation.

The residual neural network or ResNet was introduced in [63] and constitutes yet an another seminal work. It allows the training of much deeper models and has been shown to produce state-of-the-art results for various Computer Vision tasks. The most significant novelty of [63] is the introduction of a residual block where the input is directly connected to the output, and the network only learns a residual mapping. More formally, if $H(\mathbf{x})$ is the underlying mapping, a residual block would fit an alternative mapping $F(\mathbf{x}) = H(\mathbf{x}) - \mathbf{x}$, and the original mapping is recast to $F(\mathbf{x}) + \mathbf{x}$ (see Fig. 2.1). When an identity mapping is optimal, it is easier to push the residual to zero than to fit an identity mapping through non-linear layers. The ResNet architecture enabled the efficient training of networks up to 152 layers deep (compared, for example, to 22 layers for GoogleNet) since it allowed for better gradient flow and helped overcome the gradient vanishing and explosion problems.

Naturally, following its wide adaptation in various Computer Vision tasks, the ResNet architecture and the residual block where extensively used in landmark detection. A particular ResNet-based model that attracted attention due to its strong performance is

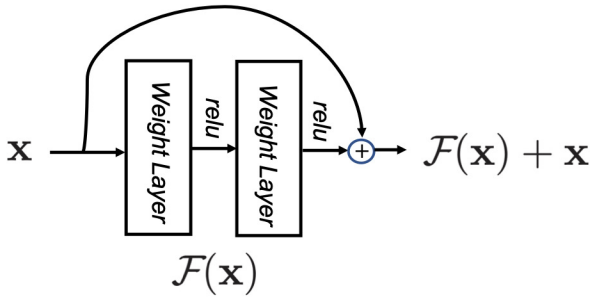


Figure 2.1: Residual block of [63]. Figure reproduced based on [63].

the Hourglass [122]. The Hourglass comprises an efficient encoder-decoder architecture that consolidates features across different scales using a single pipeline. Features are processed down to a very low resolution through max-pooling layers and then sequentially upsampled with nearest neighbours upsampling. They are also combined across scales with skip layer feature concatenation. Multiple hourglass modules can be stacked together for increased performance (referred to as Stacked Hourglass), and intermediate supervision can also be applied. An illustration of the hourglass architecture is shown in Fig. 2.2.

Variants of the Hourglass architecture have also emerged. In [34], authors added a side branch of filters to the residual unit leading to a larger receptive field. In [186] the residual unit is replaced by a multi-branch Pyramid Residual Module. PoseNet in [29] contains an hourglass-based generator and two discriminators to distinguish between reasonable and unreasonable poses. In [12], authors propose a lightweight binarised Hourglass block that yields a performance improvement while maintaining the same number of parameters. The Stacked Hourglass architecture was used for face alignment in [13]. This work will also utilise the Hourglass as a backbone for the unsupervised landmark detection as it was a state-of-the-art model during the development of this thesis.

Recently though, even stronger models have emerged and are briefly mentioned here. CPN [30] addresses hard to localise landmarks explicitly through a dual-module architecture that includes a GlobalNet and a RefineNet. The GlobalNet is trained to localise the simpler landmarks. RefineNet handles harder keypoints by considering the feature representation of GlobalNet on different scales. It is also explicitly trained on

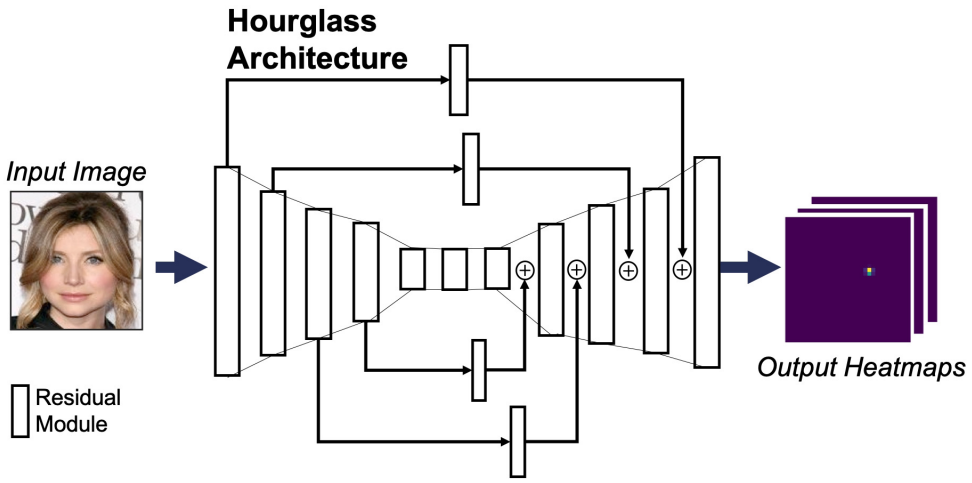


Figure 2.2: Hourglass Architecture. Each box corresponds to the residual module of [63]. Figure is reproduced based on [122].

hard keypoints through an online hard keypoint mining loss. In [180], the authors propose the Simple Baseline architecture for heatmap-based landmark detection. This Simple Baseline uses a ResNet backbone and only adds a few deconvolutional layers to generate the output heatmaps without combining features with skip layers. Despite its simplicity, this architecture results in excellent performance for both human pose estimation and tracking.

Contrary to previous encoder-decoder architectures, authors of the HRNet [155] model propose to maintain the high-resolution representation throughout the network. This is achieved through multiple connected convolutions streams that operate on different resolutions and perform input processing in parallel. HRNet and its variants [32, 66, 194, 190] show state-of-art performance, not limited to landmark detection but in a wide range of localisation tasks including semantic segmentation and object detection.

2.4 Unsupervised Landmark Detection.

The main focus of this thesis is the unsupervised detection of object landmarks. As a Computer Vision task, it was introduced relatively recently, in 2017 by Thewlis et al. [164]. Similar to its supervised counterpart, the main objective is the detection of a shape vector

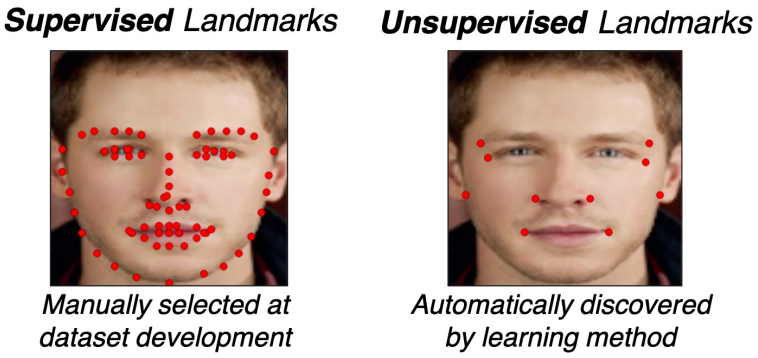


Figure 2.3: Qualitative comparison of *supervised* and *unsupervised* object locations on a facial image. *Supervised* object landmarks are manually selected to track specific target points on the object's surface. Here the common 68-point facial landmark configuration is shown. *Unsupervised* Landmark detection aims to both discover semantic landmark locations and consistently track them across instances of the same category. In this example, unsupervised landmarks are detected by our proposed approach.

$\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_K^T) \in \mathbb{R}^{2K}$ comprising of K object landmarks on instances of a specific object category. In contrast to the supervised setting, the landmark detector is learned directly from unlabelled images without manual supervision.

The objective of an unsupervised landmarks detector is to both *discover* and consistently *detect* a set of landmarks over different instances of the same object (see Fig 2.3). In that sense, unsupervised landmark detection can be considered an ill-defined problem since the target landmarks need to be both defined and consistently detected. We will refer to the set of object locations captured by an unsupervised landmark detector as *unsupervised object landmarks*. This is in contrast to *manual object landmarks* detected by supervised landmark detectors. Note that manual object landmarks are selected at the dataset development stage to capture particular object locations that are informative for downstream applications (deformable parts like the human hands and legs for human pose estimation or corners of mouth and eyes for face alignment).

Different sets of unsupervised landmarks might vary in their semantic value. For example, a set that tracks several locations on the human body but does not capture the legs or elbows would be less informative for a downstream application. In this work, our objective is to detect unsupervised landmarks with *high semantic meaning* similar to the

ones assigned by human annotators for various object categories.

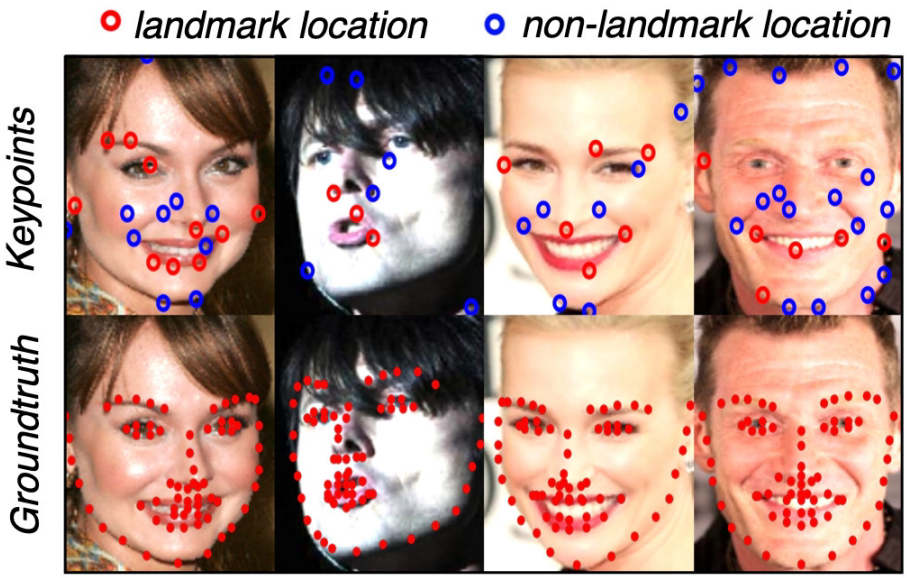
2.5 From Keypoints to Landmarks

Exploring the connection between generic keypoints and object landmarks is among the main ideas investigated in this thesis. This subsection will first introduce the reader to their various similarities, which will be further explored in later Chapters. Keypoint (or interest point) detection is the extraction of salient 2D coordinates on the image domain that can be detected consistently, with invariance to geometric or photometric transformations. Even though multiple definitions for what constitutes a generic keypoint have been proposed [76], generally keypoints can include junctions [179], corners [61], blobs [106], saliency features, locations with significant texture variations or simply points where "something occurs". Similar to unsupervised landmark discovery, detection of generic keypoints can also be considered an ill-defined problem, in the sense that there is no apriori definition of what visual structures should be detected as generic keypoints. To overcome this, interest points are commonly conditioned to have properties that are desirable for downstream applications like sparsity, invariance to viewpoint [40] or illumination changes [175], discriminativeness and repeatability [135].

This work draws an analogy between generic keypoints and unsupervised object landmarks. We identify that the aforementioned properties of generic keypoints, also make them strong candidates for unsupervised landmark locations. We find that for multiple object categories, keypoints show some consistency and *will systematically overlap with object landmarks*.

Lets we demonstrate the extent of this phenomenon on facial images, by measuring the precision[1] of generic keypoint detectors w.r.t actual ground-truth landmark locations. Results for various popular keypoint detectors are shown in in Fig 2.4 **(Table)**. For recent methods R2D2 [135] and SuperPoint [40], we observe that landmark locations

[1] As true positives, we consider keypoints within $10px$ of a landmark location (image resolution 256×256)



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

Keypoint Detector	Precision
SIFT [112]	35.2
ORB [139]	43.7
R2D2 [135]	50.7
SuperPoint [40]	**51.8**

Figure 2.4: **(Figure)**: Comparison of generic keypoints and object landmarks on facial images. Generic keypoints capture several object landmark locations (*red keypoints*) as well as non-corresponding background points (*blue keypoints*). **(Table)**: Precision of various generic keypoint detectors w.r.t 68-groundtruth landmark locations (on CelebA). SIFT and ORB that tend to detect spatially clustered points are also combined with the adaptive non-maximal suppression method of [8] to ensure a homogeneous spatial distribution of keypoints.

are consistently captured with high precision values over 50%. Visual examples of this overlap are shown in Fig. 2.4 **(Figure)**, where generic keypoints "fire" on various manually annotated landmark locations (marked with red). Based on this observation, our goal 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 also the corresponding landmark locations. We will show in later Chapters how this can be achieved through *self-training* and *correspondence recovery via clustering*.

2.6 Evaluation Metrics for Landmark Detection

Traditionally, the performance of supervised landmark detectors is quantified through the normalised point-to-point Euclidean distance between the detected landmarks $\boldsymbol{y}_{pr} \in \mathbb{R}^{2K}$ and the provided ground-truth points $\boldsymbol{y}_{gt} \in \mathbb{R}^{2K}$ on a separate test set. We will refer to this metric as Normalized Mean Error :

$$NME = \frac{1}{M} \sum_{m=1}^{M} \frac{||\boldsymbol{y}_{gt} - \boldsymbol{y}_{pr}||_2}{d} \tag{2.3}$$

where d is a normalisation constant computed for every sample. Depending on the dataset, different normalisations can be used (interocular distance for profile faces, bounding box size for datasets with large $3D$ rotations e.t.c).

For unsupervised landmark detection, the predicted unsupervised landmarks are the outcome of the learning process. As a result, manually annotated ground-truth that overlaps with unsupervised landmark locations is generally not available (even though in later Chapters we will see that our proposed approach will be able to discover supervised landmark locations). To overcome this limitation, 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 through a simple linear layer that maps the discovered landmark coordinates to those manually annotated, using a variable number of images in the training set.

More specifically, this evaluation framework first seen in [164] learns a linear regressor (a fully connected layer with parameters W_F and not bias terms) that takes as input the $2K$ unsupervised landmark coordinates $\boldsymbol{y}_{pr,uns} \in \mathbb{R}^{2K}$ and maps them to the M manually annotated landmarks $\boldsymbol{y}_{pr,sup} \in \mathbb{R}^{2M}$ as $\boldsymbol{y}_{pr,sup} = W_F \, \boldsymbol{y}_{pr,uns}$. This evaluation framework requires N annotated images from the training set for learning the W_F parameter of the linear layer, with results commonly reported for various values of N. Note that there is no backpropagation to the weights of the unsupervised landmark detector through

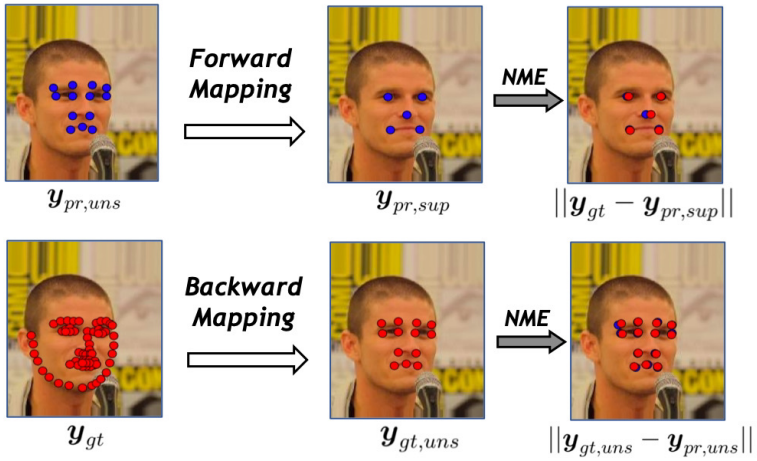


Figure 2.5: Illustration of *Forward* and *Backward* NME metrics. *Blue* landmarks are discovered by an unsupervised landmark detector and *red* landmarks are manually annotated groundtruth. *Froward-NME* maps the predicted unsupervised landmarks to the manually annotated points. For *Backward-NME*, we perform the inverse mapping from the manually annotated groundtruth to the unsupervised landmark locations discovered by the learning process. A robust landmark detector should achieve good performance for both metrics.

this training process. Originally, training augmentations were also used [164] to enhance the performance of the linear layer but were later removed [196] to minimise training complexity. Hereon, we will refer to this evaluation metric as **Forward-NME** and will be used to qualify the degree of correlation between unsupervised and manually annotated landmarks.

$$NME_{Forward} = \frac{1}{M}\sum_{m=1}^{M}\frac{||\boldsymbol{y}_{gt} - \boldsymbol{y}_{pr,sup}||_2}{d} \tag{2.4}$$

As noted in [144], **Forward-NME** only measures to which extent the discovered landmarks correlate with annotated landmarks and might dismiss the effect whereby a method can learn a few landmarks with a strong correlation with the object's geometry and keep some extra loose landmarks that are randomly placed in an image. The **Forward-NME** would not be able to show if such an effect is occurring. However, if a network has been learned to produce a set of landmarks describing the geometry of a particular object, we want all the detected points to contribute to describing the object's geometry, and have a proper method to automatically detect which learned landmarks do not correlate with it.

To measure such effect, [144] introduced an evaluation method consisting of learning a linear regressor in a *backward* manner, i.e. one that maps the *manual annotations into the discovered landmarks* or $\boldsymbol{y}_{gt,uns} = W_B\, \boldsymbol{y}_{gt}$ where W_B are the learned parameters of the linear layer. This measure, known as **Backward-NME** will be defined as :

$$NME_{Backward} = \frac{1}{K} \sum_{k=1}^{M} \frac{||\boldsymbol{y}_{gt,uns} - \boldsymbol{y}_{pr,uns}||_2}{d} \tag{2.5}$$

Backward-NME helps identify unstable landmarks. If the network contains a landmark that is randomly located, the regressor from the ground-truth points to the detected points will fail to estimate it, thus leading to large error values for the backward mapping. A visualisation of the mappings performed by the two metrics is shown in Fig. 2.5.

Moreover, we note that such a global metric does not help assess the performance of individual landmarks. It is likely that poor performance of only a subset of the detected unsupervised landmarks will bias the overall error. There are several reasons this might occur. It can be that only some landmarks are unstable, track background points or happen to be uncorrelated to the manually annotated landmarks (for example, a landmark on the ears will result in high error even if detected consistently since it is not included in the 68-facial landmark format). To address this limitation of global error metrics, we will follow standard evaluation practices of supervised landmark detection [13] and also present *CED* (Cumulative Error Distribution) curves.

2.7 Datasets

To evaluate the performance of the learned models, this thesis will present results in a wide range of datasets capturing multiple object categories. The datasets that will be examined in this work are:

CelebA-MAFL: CelebA [109] dataset of about 200K facial images annotated for 5 facial landmarks. Following standard practise, we evaluate our method on the MAFL subset, which is excluded from the training split.

AFLW: AFLW [87] contains $10,112$ training images and $2,991$ test images annotated for 21 landmarks annotated based on visibility.

LS3D: LS3D [13] is a dataset of large pose facial images constructed by annotating the images from 300W-LP [205], AFLW [87], 300VW [147], 300W [141] and FDDB [71] in a consistent manner with 68 points using the automatic method of [13]. Note that LS3D dataset is annotated with 3D points. Evaluation is performed on the LS3D-W Balanced test set that comprises 7200 including an equal number of images for yaw angles of $[0^o - 30^o]$, $[30^o - 60^o]$, $[60^o - 90^o]$.

BBCPose: BBCPose [24] 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: Human3.6M [69] is an activity dataset with a constant background containing videos of actors in multiple poses under different viewpoints. We follow the evaluation protocol of [200] 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 [200] we do not perform background subtraction to simplify landmark detection.

PennAction: PennAction [198] is a dataset of 2326 videos of humans participating in sports activities. For this experiment, we use the same 6 categories as in [110] (tennis serve, tennis forehand, baseball pitch, baseball swing, jumping jacks, golf swing). 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. The result is a

training set of 51661 images and a test set of 1776 images.

DeepFashion. DeepFashion [108] is a large dataset of clothes images annotated for 8 fashion landmarks. Since the dataset is only annotated for fashion landmarks, we infer 14 human pose landmarks using [18] and select only the subset of images depicting a full human body. This results in a train set of 12,416 images, with 500 images being kept for testing.

Cat Heads: This is a dataset of 9k images of cat heads annotated with 9 landmarks [197]. We use the test-train split of [200] with 7747 training and 1257 testing images.

CUB-200-2011: We also present qualitative results in this dataset of 11778 bird images from 200 different species [176]. We follow the same setting as [110] and remove seabird species while similarly aligning parity using eye visibility information.

When the normalised error is reported for facial datasets, similar to other methods, normalisation is performed using standard interocular distance (CelebA, AFLW, CatHeads). For human pose dataset (BBCPose and Human3.6), normalisation is performed using shoulder distance. For LS3D and PennAction we normalise with bounding box size $\sqrt{w_{bbox} * h_{bbox}}$ since interocular/shoulder distance can be very small. When calculation of the forward and backward error curves is performed w.r.t. 68 standard facial landmarks for CelebA and AFLW, these are recovered using the highly accurate method of [13] since they are not provided.

Chapter 3

Literature Review

This Chapter will provide a comprehensive review of the most relevant literature to this thesis. The main topic of this work is the discovery of an object's shape from 2D images without manual supervision. To that end, we will first discuss related unsupervised landmark detection methods, trained with either the *equivariance constraint* or through *image generation*. We will also mention methods that do not detect object landmarks explicitly but discover alternative shape representations without using manual supervision. The second part of this Chapter will focus on recent advancements in self-supervised learning and self-training, which are the main learning approaches explored in this work. Since our proposed approach utilises generic keypoints as a weak initialisation for landmark discovery, a short discussion of popular keypoint detectors will also be provided.

3.1 Landmark Discovery via Equivariance

As defined in the previous Chapter (subsection 2.4), unsupervised landmark detection refers to the computer vision task of estimating a landmark vector $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_K^T) \in \mathbb{R}^{2K}$ from unlabelled images without using any form of manual supervision. In this form, the task was first introduced in 2017 by Thewlis et al. [164]. This initial work proposed

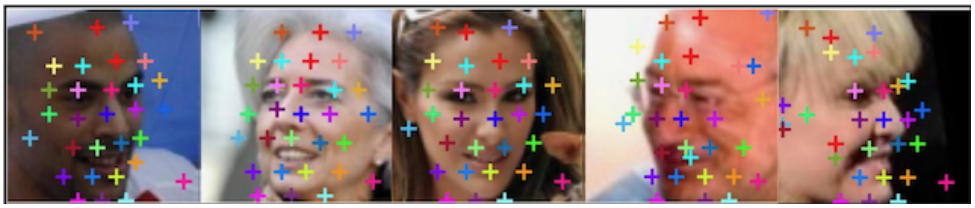


Figure 3.1: Unsupervised landmarks on facial images discovered thought equivariance. Visual results taken from [164]. Detector trained to detect 30 object landmarks.

learning a landmarks detector through equivariance to geometric transformations of an input image $\boldsymbol{x}$. Equivariance states that a mapping function $\Phi : \mathcal{X} \rightarrow \mathbb{R}^C$ is equivariant with a transformation g when:

$$\forall \boldsymbol{x} \in \mathcal{X} : \Phi(g(\boldsymbol{x})) = g(\Phi(\boldsymbol{x})) \tag{3.1}$$

This equivariance constraint can learn shape representations from unlabelled data and had been already introduced in [98] for training a generic keypoint detector. Authors in [164] extended it for unsupervised landmark discovery through the process of *viewpoint factorisation* we will discuss next. Some visual examples of unsupervised landmarks discovered from unlabelled images via equivariance are shown in Fig. 3.1.

3.1.1 Viewpoint Factorisation

Learning a landmark detector from unlabelled images with equivariance can be achieved through *viewpoint factorisation* as proposed in [164] . Authors model the objects intrinsic viewpoint-independent structure as a 3D surface S, with $S \subset \mathbb{R}^3$. Their approach is to learns a function $\boldsymbol{\Phi}_S$, to maps object points $p \in S$ to pixel q on the image surface such as $q = \boldsymbol{\Phi}_S(p, \boldsymbol{x})$. Since correspondence between object points and image pixel locations is unknown, authors propose to learning approach based on factorized viewpoints. To that end, they consider images $\boldsymbol{x}$ and $\boldsymbol{x'}$ as particular viewpoints of the same underlying surface S, and $g : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ as the warp that induced the viewpoints chance between the images such as $\boldsymbol{x'} \approx \boldsymbol{x}\, o\, g$ (occlusion not withstanding). The main idea is to factorise the

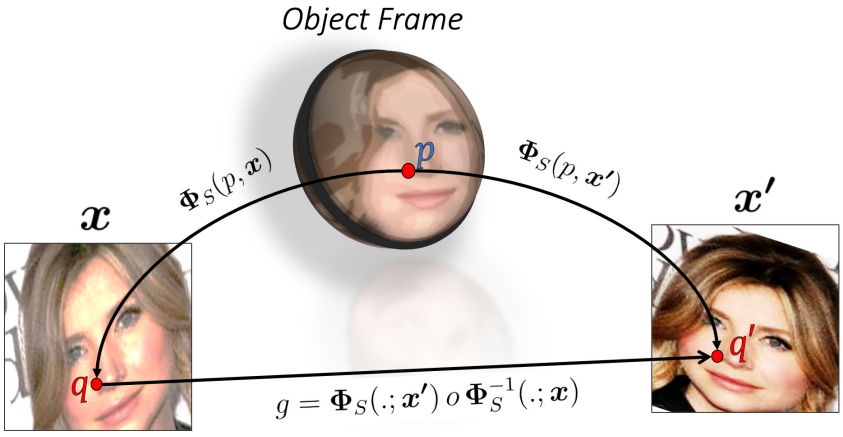


Figure 3.2: Viewpoint factorisation process. Function $\mathbf{\Phi}$ maps the point r of a $3D$ surface to the corresponding pixel q of image $\boldsymbol{x}$. $\mathbf{\Phi}$ is learned to be consistent with changes in viewpoint caused by the warp g. Figure is reproduced based on [164].

wrap g as first finding the intrinsic object point p for each pixel on image $\boldsymbol{x}$ through the inverse mapping $p = \mathbf{\Phi}_S^{-1}(q, \boldsymbol{x})$ and then finding the corresponding pixel q' in image $\boldsymbol{x'}$ as $q' = \mathbf{\Phi}_S(p, \boldsymbol{x'})$ or more formally:

$$g = \mathbf{\Phi}_S(.; \boldsymbol{x'}) \, o \, \mathbf{\Phi}_S^{-1}(.; \boldsymbol{x}) \tag{3.2}$$

Equation 3.2 can also be expressed as :

$$\forall p \in S : \mathbf{\Phi}_S(p, g(\boldsymbol{x})) = g(\mathbf{\Phi}_S(p, \boldsymbol{x})) \tag{3.3}$$

Equation 3.3 is equivalent to the *equivariance constraint* and simply states that each point on the object surface should be detected with consistency to the viewpoint induced wrap between images. Authors further extend this *equivariance constraint* to account for object deformations by introducing a common reference space S_0 called the *object frame* which ties together the possible shape variations wS of S, where $w : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ is the isomorphism that deforms S. Given a common *object frame* the *equivariance constraint* can be rewritten as :

$$\forall p \in S_0 : \mathbf{\Phi}(p, g(\boldsymbol{x})) = g(\mathbf{\Phi}(p, \boldsymbol{x})) \tag{3.4}$$

where $\mathbf{\Phi}$ learns a mapping from each pixel on the image domain to the corresponding point on the *object frame*. The whole process is also illustrated in Fig 3.2.

3.1.2 Learning a landmark detector via equivariance

Having arrived at the *equivariance constraint* through the factorisation of two views of the same object, we will next describe how this criterion can be used to train an unsupervised landmark detector. The approach followed in [164] is to sample the function $\mathbf{\Phi}$ as a set of K discrete location on the *object frame* S_0:

$$\mathbf{\Phi}(\boldsymbol{x}) = [\mathbf{\Phi}(r_1, \boldsymbol{x}), \mathbf{\Phi}(r_2, \boldsymbol{x}), \ldots, \mathbf{\Phi}(r_k, \boldsymbol{x})] \tag{3.5}$$

Since the object frame is common for different views, mapping the *same* object frame location to the corresponding pixel location on each image could be thought as detecting the p_i object landmark since $p_i = \mathbf{\Phi}(r_i, \boldsymbol{x})$ with $i \in (1, 2, \ldots, K)$.

The mapping function $\mathbf{\Phi}$ can be implemented as a deep neural network. Authors in [164] form their networks output $\mathbf{\Psi}$ as a set of K probability maps $\mathbf{\Psi}_i \in \mathbb{R}^{W \times H}$ with $i \in (1, 2, \ldots, K)$. The coordinates for each landmark can be extracted from the probability maps in a differentiatable manner using the spatial *soft argmax operation* (σ_{arg}):

$$\mathbf{\Phi}_i(\boldsymbol{x}) = \sigma_{arg}[\mathbf{\Psi}_i(\boldsymbol{x})] = \frac{\sum_u u e^{\mathbf{\Psi}_{iu}(\boldsymbol{x})}}{\sum_u e^{\mathbf{\Psi}_{iu}(\boldsymbol{x})}} \tag{3.6}$$

with $\mathbf{\Phi}_i(\boldsymbol{x}) \in \mathbb{R}^2$ being the XY coordinates of the i_{th} object landmark. The *soft argmax operation* that was first introduced in [188] essentially computes *the spatial expected value of the probability map* and, as we will discuss later in this Chapter, is regularly used

to derive fully differentiable architectures for unsupervised shape discovery.

Given the structure of $\mathbf{\Phi}$ described above, one can learn a set of object landmarks from pairs of images $\boldsymbol{x}$ and $\boldsymbol{x'} = g(\boldsymbol{x})$ of the same object through the equivariance constraint 3.4 with can be expressed as a loss term as :

$$L_{equiv} = \frac{1}{K} \sum_{i}^{K} ||\mathbf{\Phi}_i(g(\boldsymbol{x})) - g(\mathbf{\Phi}_i(\boldsymbol{x}))||^2 \tag{3.7}$$

The equivariance loss of equation 3.7 leads to a *Siamese* [85] configuration where two separate subnetworks with shared weights, process images $\boldsymbol{x}$ and $\boldsymbol{x'}$ respectively.

3.1.3 Generating image pairs for the equivariance constraint

Learning with an L_{equiv} of equation 3.7 requires a training set of triplets $(\boldsymbol{x}, \boldsymbol{x'}, g)$. In practise, even if two views of the same object are available, the viewpoint transformation g is often unknown. Instead, training triplets can be synthesized by applying a known transformation g to generate $\boldsymbol{x'}$ from $\boldsymbol{x}$. Even though $\boldsymbol{x'}$ can be a simple affine transformation of $\boldsymbol{x}$ as in [144], a common practise is to use random Thin-Plane-Splines (**TPS**) [10] wraps.

TPS applies both global and local transformations. Globally, an affine wrap (rotation, translation, scaling) is applied to the whole image and locally, a set of control points on a predefined uniform grid are spatially perturbed. In [164], authors sample 2 random TPS transformations g_1, g_2 and given an image $\boldsymbol{I}$, they form image pairs as $\boldsymbol{x} = g_1(\boldsymbol{I})$ and $\boldsymbol{x'} = g_2(\boldsymbol{x})$. The same approach is also used in other related work [163, 72]. Authors in [196] use the discovered object landmarks as an alternative set of control points. During training, they alternate between the predefined grid of controlled points and the set of discovered landmarks. On the contrary, our proposed approach will enable learning from unaligned pairs of *different images* through clustering correspondence.

Figure 3.3: Synthetic pair generation using random TPS wraps. Image samples taken from [72] with the permission of the author.

Performance can be further enhanced on video data, by sampling image pairs from nearby frames. Correspondence can be recovered in this scenario through optical flow calculation. Specifically, authors in [196] define the transformation g as the dense optical flow at 5-frame intervals, calculated using the Farneback method of OpenCV and used to form the image $\boldsymbol{x'}$ in addition to random TPS wraps. In [72] authors propose a method that can learn from nearby frames without requiring optical flow estimation. Incorporating this type of temporal supervision into the learning process can generally lead to more robust performance on video dataset, especially for more deformable object categories like the human body. Some visual examples of image pairs formed with either synthetic wraps or sampling of nearby frames (for optical flow correspondence) are shown in Fig. 3.3. Even though our proposed method shows robust performance without the use of temporal information, we will also present results with sparse optical flow supervision on video datasets.

3.1.4 Learning dense correspondence via equivariance

Motivated by [164], where the learning function $\boldsymbol{\Phi}$ is formed by sampling K discrete locations on the object frame (can be thought of as object landmarks), authors in [163] propose to use the same formulation for dense image labelling. In [163], every image location is mapped to the common object frame S_0 (compared to just K discrete points for [164]) by the mapping function $\boldsymbol{\Phi}$. The network's output is not a set of points but

a dense map $\mathbf{\Phi} \in \mathbb{R}^{W \times H \times 3}$ (one $3d$ feature for each pixel of the input image). In our work, we also learn a dense map but opt for local features of much higher dimensionality ($256d$). These $3d$ features only capture the object-specific coordinate vector (corresponding location on to S_0) for each image location. In contrast, larger features($256d$) have sufficient representational capacity to be used for landmark matching across different instances (through clustering) in our correspondence recovery framework.

3.1.5 Limitations of Equivariance

Even though equivariant detectors [164, 163] can show good performance (especially for less deformable object categories like human or animal faces), they also demonstrate various drawbacks. One noticeable limitation is that the equivariance constraint is insufficient to learn a landmark detector since training can degrade to a trivial solution where all image pixels are always mapped to a constant point on the reference frame. In [164], authors avoid degenerate solution through the following *diversity loss*:

$$L_{div} = \frac{1}{K^2} \sum_{r=1}^{K} \sum_{r'=1}^{K} \sum_{u} p(u|\boldsymbol{x}, r) p(u|\boldsymbol{x}, r') \tag{3.8}$$

This term is lower when the support of the probability maps for different landmarks is disjoint or has minimum overlap. We note that this loss assumes that different landmarks mostly appear on different locations on the image domain. Even though this assumption might hold for some object categories (for example, frontal facial images), it might degrade performance for objects that appear under large 3D rotations or more deformable objects where different landmark locations can regularly overlap (for example, facial landmarks tend to appear in very close proximity on profile facial images). For dense image labelling, trivial solutions can be avoided by learning labels that are not only equivariant with geometric transformation but are also distinctive in the sense that for each pixel pair

(u, u'):

$$\mathbf{\Phi}(\boldsymbol{x}, u) = \mathbf{\Phi}(g\boldsymbol{x}, u') \iff gu = u' \tag{3.9}$$

Equation 3.9 can be encoded on *distinctiveness loss* that similarly to the *diversity loss* mentioned previously is used along equivariance to learn a robust shape representation.

Another weakness of equivariant detectors is that consistency across different samples is not explicitly enforced. As mentioned, the detector learns a mapping to common object frame S_0 with equivariance to geometric transformations of the *same* image. The constraint does not explicitly guarantee that S_0 will remain consistent for different samples (for example, faces with different identities). This means that the same location on the object frame can potentially correspond to the eye in one facial image and the mouth corner in another. In practice, authors of [164] find that inter-class consistency occurs to some degree due to the generalisation of the learned algorithm.

One work that tries to address this limitation is [162]. The authors discuss the relationship between landmarks and local descriptors. Landmark locations can be through as tiny $2D$ descriptors with enough representational capacity to express the landmark's index. On the other hand, high dimensional embedding vectors can represent instance-specific details for each object location. The proposed method learns local high-dimensional embeddings that are invariant to intra-class variations (like change in the identity of the object instance), thus leading to semantic consistency across different instances. This is achieved through a technique that is deemed *Descriptor Vector Exchange*, where the embeddings extracted from an image $\boldsymbol{x}$ are learned to be exchangeable with the ones extracted from a separate auxiliary image of a different instance $\boldsymbol{x_a}$. Our proposed framework also learns local embedding that are invariant to intra-class variations. Features are not exchanged between the source and auxiliary images like in [162], but invariance is achieved by learning them to be close in feature space through a contrastive loss.

Finally, landmarks discovered through equivariance are simply the outcome of the learning process and are not explicitly biased to capture object locations with semantic meaning. In the next section, we will discuss how later work attempts to enhance the semantic meaning of the discovered landmarks by combining unsupervised landmark detection with image generation or reconstruction. In contrast, the framework proposed in this thesis achieves the discovery of highly semantic object landmarks through bootstrapping from generic keypoints.

3.1.6 Equivariance vs Invariance

So far, we have discussed various ways under which shape can be learned without manual supervision using a criterion based on equivariance to geometric transformations. Interestingly, authors in [33] suggest that invariance to geometric transformations can be sufficient to learn object landmark representations as well. Given the definition of equivariance $\Phi(g(\boldsymbol{x})) = g(\Phi(\boldsymbol{x}))$, invariance can be considered a special case of equivariance where g is the identity mapping or $\Phi(g(\boldsymbol{x})) = \Phi(\boldsymbol{x})$. As we will discuss later in this Chapter (Subsection 3.4), a mapping function Φ can be trained to be invariant to geometric transformations through self-supervised learning and particularly contrastive learning. Contrastive learning aims to learn representations that are further apart in embedding space for different samples but similar for augmented views of the same object, thus encoding invariance to geometric and photometric transformations.

Early layers in DNNs tend to learn representations that are equivariant to transformations, with invariance only gradually emerging as we ascend the layers. This is a well-known phenomenon observed in various works [97, 192, 1]. Motivated by this observation, authors in [33] propose to export a shape representation from the intermediate layers of a ResNet trained with the MOCO framework [63] (discussed in Subsection 3.4), since these layers would offer the best trade-off between equivariance and invariance. This representation is formed as a hypercolumn that first interpolates features from intermediate layers and concatenates them to produce a dense representation that can be mapped to

object landmark locations using only a few annotated samples.

3.2 Landmark Discovery via Generative Modeling

To address the lack of semantic meaning of object landmarks discovered via equivariance, a more recent line of work [196, 72, 144] attempts to combine unsupervised landmark detection with image generation or reconstruction. The main idea explored by these methods is that landmark locations that are informative for image generation would also have increased semantic meaning. Commonly, a fully differentiable framework is proposed where the object's shape is first distilled to object landmark coordinates/heatmaps that are subsequently used to condition a reconstruction/generation task. The learning algorithm then discovers the object locations that achieve the strongest reconstruction/generation result.

3.2.1 Landmark Discovery via Reconstruction

One of the first methods that explored this learning approach was Zhang et al. in [196]. The proposed architecture comprises two parallel encoders to export feature and shape representations and a decoder to reconstruct the original image. More specifically the encoder includes a feature extractor $F(\boldsymbol{x};\theta_F) \in \mathbb{R}^{W\times H\times d}$ and a landmark detector $L(\boldsymbol{x};\theta_L) \in \mathbb{R}^{2K}$ where θ_F, θ_L are the model parameters. The landmark detection stream first extracts a set of $K+1$ detection score maps (one channel represents the background). These maps are transformed to probability estimates through a softmax layer, and the coordinates (x, y) of the maximum response for each map are estimated using the *soft argmax operation* operation as described in equation 3.6 resulting in

$$[x_1, y_1, \ldots, x_k, y_K]^T = L(\boldsymbol{x};\theta_L) \tag{3.10}$$

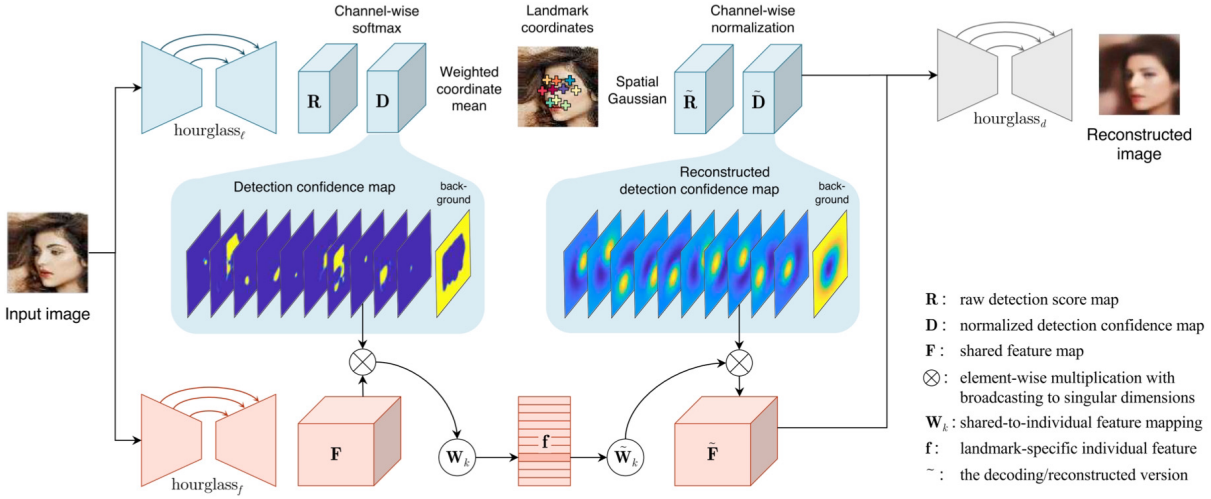


Figure 3.4: Unsupervised landmark detection framework of [196]. A landmark detection stream extracts $K+1$ raw score maps that are transformed into confidence maps by placing a gaussian on the estimated maximum response (x, y) coordinate. The confidence maps are used to sample a set of feature vectors $\boldsymbol{f}$ on the corresponding landmark locations that are fed into a decoder stream to reconstruct the input image. Figure is from [196] with the permission of the author.

Then a descriptor is sampled for each landmark location through max pooling weighted by a soft mask centered on the corresponding landmark location (x_i, y_i) of the i_{th} landmark. The features for the K object landmarks (plus a background feature) are concatenated to a feature vector $\boldsymbol{f} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_K, \boldsymbol{f}_{K+1}]$ that is the input of a decoder module that reconstructs the image. An illustration of the proposed pipeline is shown in Fig. 3.5.

To encode desirable properties to the discovered object landmark location, multiple soft constraints are enforced. These are a *concentration constraint* to concatenate the response of the detection maps to a local region, a *separation constraint* to encourage uniform distribution of discovered landmarks on the object's surface as well as the standard *equivariance constraint* (that is necessary for robust performance as found in [72]). Landmarks detected by [196] follow a rather grid-like layout, mainly due to the use of the separation constraint (as it can be seen in Fig. 1.3 of the introduction Chapter). Moreover, we find that contrary to our work, this approach does not scale well to complex object deformations and out-of-plane rotations (also noted in [111]).

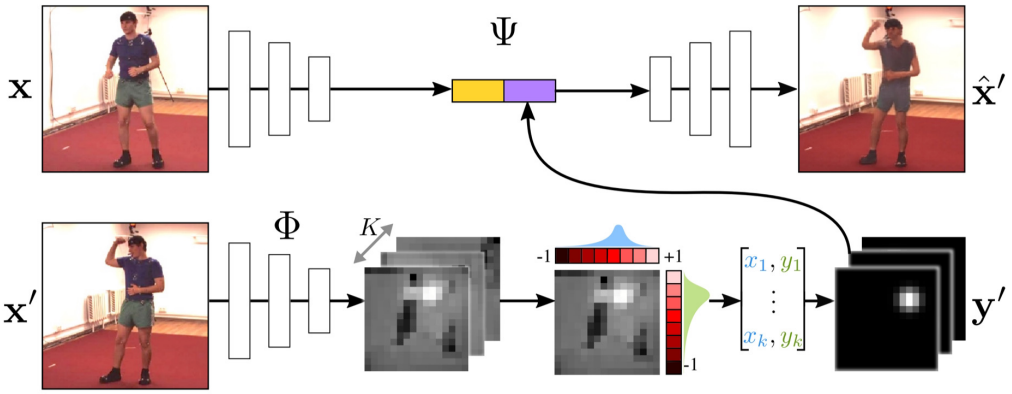


Figure 3.5: Unsupervised landmark detection framework of [72]. A landmark detection subnetwork distills the shape of object in image $\boldsymbol{x'}$ as a set of K heatmaps with $2D$ gaussians centered on the landamark coordinates. These heatmaps are stacked along with appearance features from image $\boldsymbol{x}$ to generate $\boldsymbol{x'}$. Figure is from [72] with the permission of the author.

3.2.2 Landmark Discovery via Generation

A related framework that attracted much attention due to its simplicity and improved performance is the landmark detection approach of Jakab et al. in [72]. Instead of auto-encoding the input image as in [196], authors generate an image that combines the appearance of an object as depicted in $\boldsymbol{x}$ and the shape of the object as depicted in $\boldsymbol{x'}$.

More specifically, the authors aim to learn a function Φ that extracts the shape of an object in an unsupervised manner in the form of K probability maps. To that end, they consider two images $\boldsymbol{x}$ and $\boldsymbol{x'}$ of the same object under a different shape configuration. These can be synthetically generated using TPS or sampled as nearby frames on video data. The learning algorithm aims to generate $\boldsymbol{x'}$ as $\bar{\boldsymbol{x}}' = \Psi(\boldsymbol{x}, \Phi(\boldsymbol{x'}))$. To avoid trivial solutions where Φ is learned to be the identity mapping $\Phi(\boldsymbol{x'}) = \boldsymbol{x'}$, authors enforce a tight bottleneck to the output of Φ where raw output heatmaps are transformed to probability maps similar to [196] as described in the previous subsection. Both Φ, Ψ are learned jointly using a perceptual loss. The loss is defined as :

$$L_{percept} = \sum_{l} a_l ||\Gamma_l(\boldsymbol{x'}) - \Gamma_l(\bar{\boldsymbol{x}}')||^2 \tag{3.11}$$

where Γ_l is the output of an off-the-shelf pre-trained network at layer l and a_l are weights for different layers. An illustration of the overall framework can be seen in Fig. 3.4. This approach is simpler than [196] since extra constraints (including equivariance) are not required for good performance. Moreover, this method can learn directly from video data since they are a natural source of images with the same appearance and varying object geometry without need for external components (optical flow estimation algorithms).

The framework of Jakab et al. [72] influenced subsequent works. In [91] it is modified to learn shape representations that are informative for reinforcement learning and control, in [150] is used for the task of animation of object categories where landmark annotations are not available, and in [148] for video interpolation and prediction. In [144], authors propose to extend [72] for learning object landmarks through unsupervised domain adaptation. The main idea of this work (as also shown in Fig. 3.6) is to alter the structure of the Φ module that now comprises a pre-trained landmark detector with weights that remain frozen during training and a projection matrix to adapt into new object categories. More specifically, an hourglass $\Phi(;\theta)$ is first trained with standard heatmap regression for an object category where manual annotations are available. Then, the weights of of the L_{th} convolutional layer of Φ are re-parametrized as $\theta' = \phi(\boldsymbol{W}_L, \theta_L)$ where $\boldsymbol{W}_L$ is a projection matrix and ϕ a linear function. During training, only the projection matrix $\boldsymbol{W}_L$ is learned while the weights θ remain frozen, leading to a more constraint optimisation problem with fewer learnable parameters (reduction by a factor of 9). The method of [144] is orthogonal to our approach since it can be applied as a drop-in replacement to our learning framework, although it was not examined for the work presented in this thesis.

Another recent extension of the landmark detection framework of Jakab et al. [72] is [103]. Authors in [103] highlight the limitation of learning from augmented pairs of the same underlying image. They extend the framework of [72] through a two-path training framework that aims to improve intra-subject consistency through the fusion of auxiliary representations from unpaired images. Our proposed method also learns from unpaired images through deep clustering for correspondence recovery.

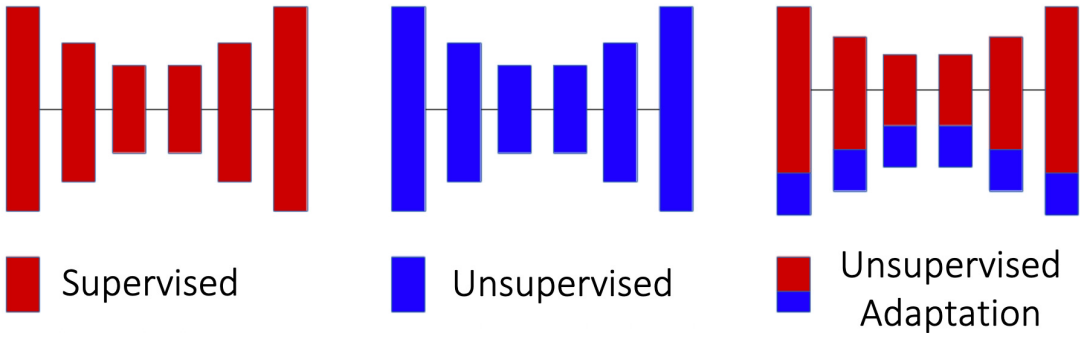


Figure 3.6: Unsupervised landmark discovery thought domain adaptation as proposed in [144]. Knowledge learned from supervised training on a separate object category can be transferred to a novel by keeping the pretrained network frozen and training only a small number of parameters. Figure is from [144] with the permission of the author.

3.3 Discovery of Alternative Shape Representations

Even though this thesis focuses on the unsupervised discovery of $2D$ object landmarks on the image domain, alternative shape representations are also regularly explored in recent literature. One such example we have already discussed is dense image labelling (see Subsection 3.1.4), where a $3D$ label is recovered for *every pixel* of the input image. In this section, we will mention recent unsupervised methods that, contrary to ours, do not extract object landmarks but related shape representations.

3.3.1 Learning object symmetries via equivariance

In [165], authors follow the same framework used for dense image labelling in [163] to model object symmetries. Authors consider object categories with bilateral symmetry and learn a mapping to a dense object frame S_0 that is now able to capture the symmetry of natural objects. This can be achieved with minimal extension to the learning formulation of [163], simply by randomly applying the reflection operation on both the image and the embedding space. In this thesis, we follow a related approach. However, in contrast to [165] that aims to learn dense object symmetries, our method recovers landmark correspondences under image reflection (through deep clustering). This allows us to include flipping augmentation during training (commonly not utilised for unsupervised landmark detection since landmark

correspondence is unknown) and learn a more robust landmark detector.

3.3.2 Unsupervised Discovery of 3D Keypoints

Related to $2D$ landmarks detection is also the objective of [159], where 3-dimensional landmarks ($2D$ points with depth) are learned without manual supervision. To enforce the discovery of semantically useful object locations, an end-to-end framework is proposed where optimisation is performed with respect to two downstream tasks. These are the recovery of relative pose between 2 views and consistency across $3D$ object transformations. Contrary to our approach, where landmarks are learned from raw images, this work requires paired multiview images that are rendered from ShapeNet [23], a large database of $3D$ models of various objects.

3.3.3 Unsupervised Disentanglement of Shape and Appearance

So far, we have discussed various approaches for discovering object landmarks through generative modelling [72, 196]. Related to this research direction is also a set of methods that learn shape representations through the disentanglement of shape and appearance. In [149] authors propose Deforming Autoencoders, a framework that distils a separate latent embedding for shape and appearance and then combines them to reconstruct the original image. Shape is represented through *deformation modelling* comprising both global and local components. The global component is implemented as a Spatial Transformer layer [70] that synthesises deformation fields as an expansion on a fixed basis and local components are the spatial gradients of local warping fields (displacement of subsequent pixels). The aforementioned shape representation can be informative of object landmark locations but can also be used for manipulating the pose of an input image.

Similarly, in [111] shape and appearance are also disentangled. This work extracts a part based shape representation where the i_{th} part is formed as a part activation map

$\sigma_i(x) \in \mathbb{R}^{W \times H}$. The whole framework is learned from augmented pairs of the same by integrating both invariance and equivariance constraints into an image reconstruction task. Authors present comparisons with landmark detection methods by using the part mean as their landmark location estimate.

3.3.4 Learning Object Landmarks from Unaligned Data

A different perspective for object landmark detection is explored in [73] where authors address the problem from the perspective of Image Translation. The proposed method (that is later extended in [74]) uses the CycleGAN [204] to translate the input image into a skeleton image through a skeleton bottleneck, implemented as an analytical differentiable renderer that removes appearance leakage. The model is trained end-to-end for image generation (similar to [72]), as well as through an adversarial loss, where a discriminator aligns the distributions of generated skeletons to genuine skeletons. Contrary to our method, this approach requires prior information in the form of unaligned pose data (which can be provided from mocap datasets) that are only available for limited object categories.

3.4 Self-supervised Learning

Our proposed framework for unsupervised landmark detection exploits recent advancements in Self-Supervised Learning **(SSL)**, and this section will provide a brief introduction to this emerging learning approach. For a more detailed description, the reader can refer to recent surveys [77, 107].

3.4.1 Motivation of Self-supervised Learning

SSL has recently attracted significant attention, given its potential for alleviating the need for expensive manual annotations as well as addressing other limitations of supervised

learning like generalisation error [52, 172], spurious correlations [171, 107] and lack of robustness to adversarial examples [161]. The main idea of SSL is that supervision is obtained by the data itself through a specially designed pretext task. This task can be based on a semi-automatic process or the prediction of part of the data from other parts. Following [107], mainstream SSL can be summarised into three main categories.

Generative: Input x is first mapped to a representation z by an encoder and a decoder reconstructs x from z (or parts of z). Common pretext tasks that follow this approach are image reconstruction [42, 83] for Computer Vision, word prediction (next word prediction [131], next sentence prediction [41], sentence order prediction [93]) for Natural Launguage Processing or link prediction [59, 84] for graph learning.

Adversarial: An encoder is trained to generate samples, and a discriminator is trained to distinguish between real and fake samples. Here, input representation z is modelled implicitly, and learning is facilitated through a distribution divergence loss like Wasserstein Distance or JS-divergence. The most prominent example of this approach are GANs [118] and related variants [44, 45], but it is also utilised for the development of various popular pretext tasks like image colorization [94], inpainting [128], super-resolution [95] and others.

Contrastive: Input x is mapped to a representation z by an encoder. The representation is learned to be informative for comparing with other representations by predicting a learning target. The learning target can be defined as a data representation learned by the network itself and varies on-the-fly during training.

In this thesis, we will learn landmark representations through Contrastive SSL. A more detailed discussion of recent contrastive methods is provided in the following subsection.

3.4.2 Contrastive Self-supervised Learning

Contrastive SSL can be divided into *context-instance* and *instance-instance* contrast [107]. *Context-Instance* contrast learns by associating representations between local and global

features of the input sample. Popular examples are methods that explore the spatial relationships between image patches, for example, the solving of jigsaw puzzles [81, 119] or geometric transformation like image rotation [53]. Other methods aim to maximise the Mutual Information between local and global context. Examples are Deep InfoMax [64] for visual feature learning, Contrastive Predictive Coding [173] for speech recognition and InfoWord [86] for sentence representation.

Intuitively, one drawback of *context-instance* contrast is that the pretext task used to learn representations (for example, arranging the patches of the jigsaw puzzle) is very different compared to commonly used downstream tasks (classification, localisation, e.t.c). As a result, a large part of the learned network has to tune into the pretext task's particular details. As an alternative, *instance-instance* contrast learns through the relationships of different inputs on the instance level by grouping similar samples closer in feature space and pushing diverse samples further apart. One way to form a *instance-instance* pretext task so it resembles supervised learning is through cluster discrimination [185, 104, 181, 19, 125, 102, 207, 184, 75, 6, 20]. Commonly, these approaches group the samples into different clusters based on the learned representations and a CNN is trained either to recognize samples belonging to the same cluster [102], by using the cluster assignments as pseudo-labels [125, 19] or by casting clustering-assignment to an instance of the optimal transport problem [6, 20].

Recently, an even stronger performance has been achieved by instance discrimination methods like MoCo [62] and SimCLR [25]. These approaches achieve state-of-the-art performance by forming positive pairs as different augmentations of the same underlying image and discriminating between positive and negative pairs (views from different images). Performance increase for these kinds of methods was enabled through several technical advancements. The number of negative samples was increased substantially either through momentum contrast [62], larger batch sizes[25] or the use of memory banks [178]. The importance of hard positive mining is also explored. In SimCLR, authors use strong data augmentations to form positive pairs. In InfoMin [167], positive pairs are formed as views

with minimum mutual information between them. Finally, even more recently, a line of work [58, 27] discarding negative samples altogether, for example, through learning a representation target that is the exponential moving average of the learned encoder. In this work, motivated by cluster and instance discrimination to develop a novel pretext task for learning landmark correspondence. To our knowledge, this is a novel perspective proposed for the first time in this work.

3.4.3 Self-Training

Another powerful learning approach that will be used in this thesis for utilising unlabelled data is *Self-training*. Self-training or pseudo-labelling refers to a set of methods where a model's own predictions are used for model training. It can be applied as either a semi-supervised or self-supervised learning approach. In a semi-supervised learning context, a model is first trained on a small labelled dataset and subsequently applied to unlabelled data. For self-supervised learning, labelled data are not available. A notable example of self-training as a flavour of self-supervision is cluster discrimination (subsection 3.4.2), where pseudo-labels are formed as clustering assignments that are used to train a new model iteratively.

Commonly in self-training, predictions are converted to pseudo-labels when detected with high confidence [96, 153, 182] but can also formed through model [123] or transformation ensembles [132], selected by applying curriculum learning principles [22] or retained only when model uncertainly is low [136]. Strong regularisation during training [182] is also used to limit overfitting of noisy pseudo-labels. Recently, [183] demonstrated very strong performance for ImageNet classification following an iterative self-training paradigm where a teacher network produces pseudo-labels to train student network that is noised in the form of dropout, randaugment [38] and stochastic depth [65]. Evidence suggests [208] that the improvements from self-supervised pretraining are orthogonal to semi-supervised self-training. Authors in [26] combine the two and manage to surpass supervised performance for Imagenet classification using only 1% or 10% percent of labels.

Most self-training approaches focus on the task of image classification. More similar to our approach are methods for unsupervised segmentation [39, 80], foreground-background segmentation [48, 154] and salience object detection [193]. In this work, we propose an iterative self-training framework for landmark detection bootstrapped by generic keypoints. To our knowledge, there is no other framework based on self-training for unsupervised object landmark detection.

3.5 Detection of generic keypoints

Keypoint detection is a critical step for any sparse image matching algorithm, for example, Structure-from-Motion [2], Simultaneous Localisation and Mapping (SLAM) [14], 3D reconstruction, Photogrammetry and others. To our knowledge, exploring generic keypoints as a weak initialisation for landmark discovery is a novel perspective first explored in this work. Given the significance of the task, there is long and extensive literature on interest point detectors. Some popular approaches will be mentioned here, but for a comprehensive review, the reader can refer to [76]. Before deep learning based solutions became the norm for most computer vision tasks, a wide range of handcrafted keypoint detectors was developed. Harris [61] method, proposed in 1988, detected keypoints based on first-order intensity variations and became widely popular. Following Harris, researchers explored various criteria for keypoint detection like self-dissimilarity [152], curvature-based methods [120] and blob detection in multi-scale space [106]. One of the most influential works in the subject is the scale-invariant feature transform or SIFT [112] based on second-order intensity variations. SIFT was proposed in 2004 and was followed by several variants [9, 3, 116, 100, 117, 4, 143]. Moreover, machine learning based keypoint detectors also emerged, most notably the FAST [138] detector that was later refined by subsequent works like ORB [139] and BRISK [99].

More recently, deep learning-based keypoint detection approaches have resulted in a very strong performance. Authors of LIFT [188] trained a DNN for keypoint detection using

Figure 3.7: Visual examples of generic keypoints detected by *SuperPoint* [40] on facial images.

image patches under different ambient conditions. In [199], authors use features produced by the handcrafted method TILDE [175] as pseudo-labels to train a deep keypoint detector, and in [98] generic keypoints are learned through a covariance constraint. Quad-networks in [145] propose a training framework based on ranking keypoint robustness under affine transformation. Recently, authors in [40] proposed to train a keypoint detection through self-training and synthetic pretraining. In this approach, deemed *SuperPoint*, a detector is first pre-trained on 2D geometrical shapes (quadrilaterals, triangles, lines and ellipses) that are automatically rendered and labelled for keypoints, on junctions, ellipse-centres and end-of-line segments. This pre-trained detector is used to bootstrap an iterative self-training framework on MS-COCO unlabelled images. Performance is further improved through a training augmentation where random homographies are applied on wrapped copies of the input image, a technique the authors refer to as Homographic Adaptation. *SuperPoint* is suitable for a large number of multiple-view geometry problems and will be the primary keypoint detector used in this work to bootstrap our landmark detection framework. Some visual examples of generic keypoints detected by *SuperPoint* on facial images can be seen in Fig. 3.7.

Chapter 4

From Generic Keypoints to Object Landmarks via Self-Training.

This work explores two novel perspectives for the Unsupervised Discovery of Object Landmarks, (1) *Self-Training on Generic Keypoints* and (2) *Clustering Correspondence*. To introduce our investigation of these two novel ideas, we will first focus on the former. To that end, a novel approach for unsupervised discovery of landmarks via *iterative self-training on generic keypoints* will be introduced. At this stage, correspondence will not be recovered via deep clustering. Instead, a separate step will be used where detected landmarks are aligned to a single point template. Even though this approach will be less flexible compared to the work presented later in this thesis, we will see that it can result in good performance and the detection of semantically meaningful landmarks in various datasets. The ideas of this Chapter would be the foundation of the methodology developed throughout the rest of this thesis.

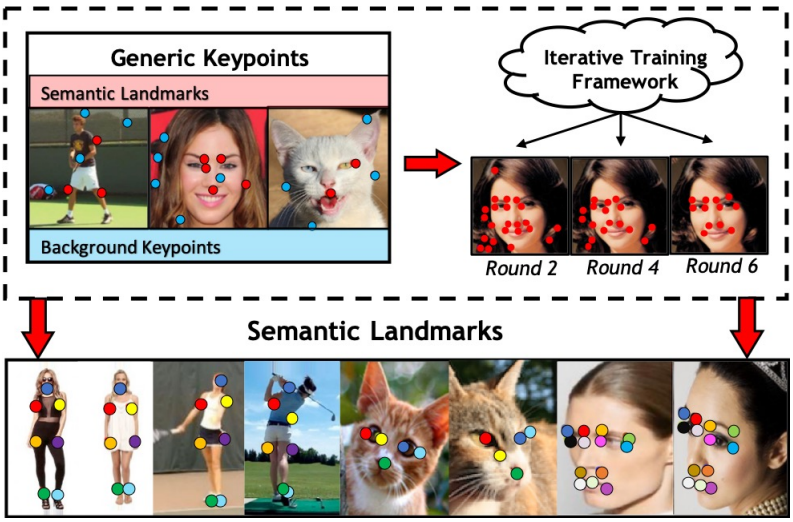


Figure 4.1: Across various object categories, object landmarks can be captured as generic keypoints. Generic *"SIFT-like"* keypoints can be considered a noisy mixture of object landmark locations (*red points*) along with several points in the background and non-corresponding points on the objects of interest (*blue points*). Our proposed self-training framework can progressively filters out noisy keypoints and converge to semantic landmark detection.

4.1 Motivation

As noted previously (Section 2.5), generic keypoints share several similar properties to object landmarks, like sparsity, invariance to viewpoint [40] or illumination changes [175], discriminativeness and repeatability [135] to just name a few. For many object categories, manually annotated landmarks are mostly located on edges or corners on an object's surface, and as such, they can be detected by a generic keypoint detector. Our assumption is that detectors showing good repeatability will be able to *capture the same landmarks on several different training images depicting objects with similar appearance and pose* (see also Subsection 2.5). In a sense, we think of detected keypoints as a *noisy mixture* containing object landmarks as well as points in the background and non-corresponding points on the objects of interest (Fig. 4.1).

Given the large percentage of noisy points it contains, attempting to learn through such a mixture could be considered impractical. Classification literature would suggest that high levels of label noise would have severely degrading effects on accuracy [206]. Recently though, it has been shown that Deep Neural Networks *(DNNs)* can generalise

well even when trained with *extreme label noise* (given that the nature of the noise is random). Authors in [5] note that despite their ability to memorise random noise, deep networks tend to learn easier (non-random) patterns in early training stages. Moreover, contrary to traditional thinking, models with higher capacity can fit noisy examples in a way that does not interfere with learning from clean data, and explicit regularisation can decrease noise overfitting without hindering the ability to learn.

The ability of a DNN to filter our noisy points can also be enhanced given appropriate hyperparameter settings like the use of lower learning rates and larger batch sizes. [137] showed that, for the task of classification, using larger batch sizes and lower learning rates can significantly increase DNN's robustness to label noise. Authors speculate that noise levels are shown to reduce the effective batch size as random influences on gradient updates can be thought to cancel out. Also, reduced learning rates result in smaller, more stable steps towards an inherently noisy gradient direction. In this Chapter, motivated by recent advances in learning under label noise, we will explore relevant techniques and appropriate hyperparameter settings (early stopping, smaller learning rates, larger batch sizes and explicit regularisation) to train robust landmark detectors through a noisy initialisation.

Our proposed *self-training* landmark detection framework does not require manual supervision and is bootstrapped by generic keypoints. The SuperPoint detector [40] (see Subsection 2.5) is initially applied on the training set to produce a pseudo-groundtruth. We use this pseudo-groundtruth to train a deep landmark detector. Once the network is trained, it is iteratively applied to the training set to produce, on each iterative round, an "improved" pseudo-ground-truth. We show that with appropriate setting of hyperparameters like batch size, learning rate, explicit regularisation and early stopping, after a few iterations of our iterative self-training framework, our detector is able to filter out background keypoints and converge to semantic object landmark locations.

Our detector is initially trained to only detect the landmark locations (similar to SuperPoint) without point correspondences between different instances of the same category. In a subsequent step, we will recover correspondence for the improved pseudo-groundtruth

by aligning the detected landmark locations with a single point template and using it to train our final model for landmark detection. Since noisy mixture of generic keypoints also includes manual object landmark locations, our approach will be able to recover highly semantic landmarks while maintaining similar accuracy to recent unsupervised landmark detectors. In the next Chapter, we will present a more flexible approach for performing correspondence recovery alongside self-training that will result in a more robust landmark detection framework.

In summary, the **contributions** of this Chapter are:

- We propose an unsupervised landmark detector that can identify semantic object landmarks. We are able to detect 13 of the 68 landmarks detected by supervised facial landmark detectors and 7 of the 16 landmarks detected by supervised human pose estimation networks.
- We explore a number of techniques for landmark localisation in the presence of extreme noise and show that with appropriate setting of hyper-parameters like regularisation, batch size, learning rate, and early stopping, our detector, after a few iterations of the iterative self-training framework, is able to filter out background keypoints and converge to meaningful landmarks.

4.2 Method

4.2.1 Problem statement

We will start by defining the problem of unsupervised landmark detection, bootstrapped by generic keypoints. Let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^{W \times H \times 3}\}$ be a set of N images corresponding to 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 discover. Note that the structure of $\mathcal{Y}$ implies that both order and landmark correspondence is recovered. Also, note that K is the underlying number of "discoverable" object landmarks which our method discovers. Here K will be the number of points on the single template that will be used for correspondence recovery (see Subsection 4.2.4).

4.2.2 Iterative training framework.

An illustration of our proposed iterative training framework can be seen in Fig. 4.2. For the first stage of our approach, generic keypoints produced by a deep keypoint detector (SuperPoint [40]) are exploited as pseudo labels for landmark localisation. A new network is trained over every iterative round using pseudo labels inferred initially from SuperPoint and then onwards from the detector trained in the previous round (Section 4.2.2). Throughout training, we utilise appropriate setting of training hyperparameters to leverage the generalisation abilities of deep neural networks when trained with noisy data (Section 4.2.3). In Section 4.2.4 we demonstrate how correspondence can be recovered through alignment with a single point template and used to train a deep landmark detector iteratively.

A deep neural network learns the mapping $\mathbf{\Psi}_\theta : \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X} \in \mathcal{R}^{H \times W \times 3}$ is the space of RGB images and $\mathcal{Y} \in \mathcal{R}^{H_o \times W_o \times 1}$ is the space of single-channel confidence maps. At every training round t, our method learns a new model $\mathbf{\Psi}_\theta^t$ from scratch using pseudo-groundtruth detected by $\mathbf{\Psi}_\theta^{t-1}$ (learned at the previous training round). Initially, for $t = 0$, pseudo-ground truth is detected by SuperPoint, the deep generic keypoint detector of [40]

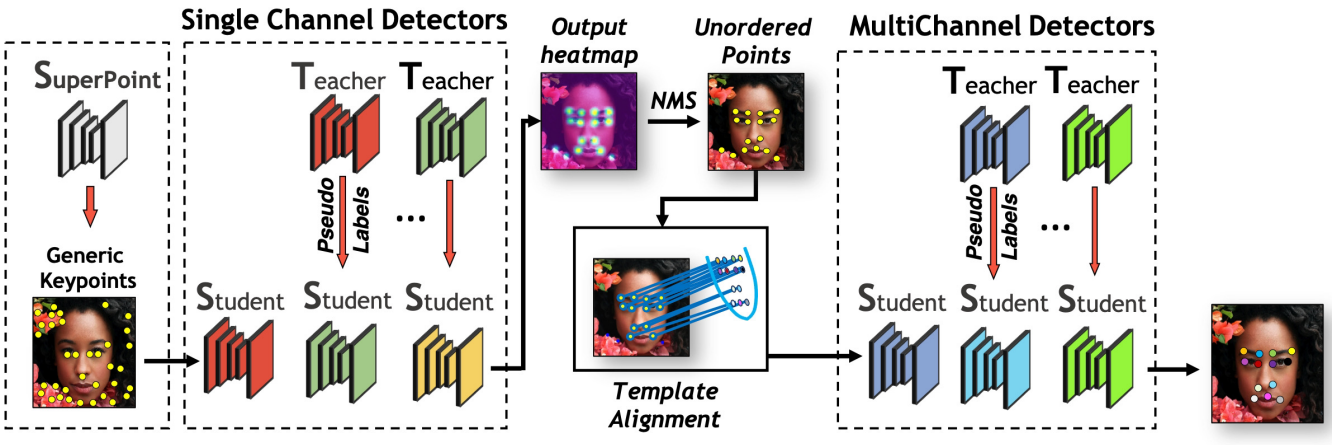


Figure 4.2: Using the output of Superpoint as initial pseudo labels, we iteratively train a network for landmark localisation whose output progressively produces improved pseudo-labels for the next round. Landmark correspondence is recovered by alignment with a single template and used to train a multichannel landmark detector iteratively.

that is also trained without manual supervision. Note that a set of output heatmaps is produced under the typical supervised setting, one for each landmark. Because we start from a generic keypoint detector, namely SuperPoint, there is *no landmark correspondence* in the provided pseudo-groundtruth, and hence our method outputs initially only a single heatmap that learns to infer multiple keypoints.

During training, a heatmap $\mathbf{H} \in \mathcal{Y}$ is produced for each image j in the training set, and multiple landmark locations are extracted through thresholding and Non-Maximum Suppression (NMS) as in [40]. NMS is applied to suppress all the keypoints that are close to the maximum keypoint in a local neighbourhood. Specifically, the $[\hat{x}, \hat{y}]$ coordinates of the i^{th} keypoint is extracted from $\mathbf{H}$, by identifying the pixel with the maximum detection confidence within a window of size $\sigma \times \sigma$. We can express this operation as :

$$[\hat{x}, \hat{y}] = \underset{x,y}{\operatorname{argmax}} \{\mathbf{H}(x,y) \mid 0 \leq x, y < \sigma\} \quad st.\ \mathbf{H}(\hat{x}, \hat{y}) > t \tag{4.1}$$

where t is a minimum detection threshold. At this stage of our algorithm, NMS operates a filtering mechanism that discards low-confident, spatially clustered keypoints from $\mathbf{p}^j$.

Algorithm 1: Proposed Self-Training Framework

Require: Image set $\mathcal{X}$
Require: Keypoint detector $SuperPoint_{Net}$
Require: Point template $\mathcal{T}$

$pseudoAnnot = SuperPoint_{Net}(\mathcal{X})$;
for $round \leftarrow 0$ **to** N **do**
  $Net = Model(Channels = 1)$;
  $Train(Net, \mathcal{X}, pseudoAnnot)$;
  $pseudoAnnot = NMS(Net(\mathcal{X}))$;
end
$pseudoAnnot_{correspdnce} = Align(NMS(Net(\mathcal{X}), \mathcal{T})$;
for $round \leftarrow 0$ **to** M **do**
  $Net = Model(Channels = len(\mathcal{T}))$;
  $Train(Net, \mathcal{X}, pseudoAnnot_{correspdnce})$;
  $pseudoAnnot_{correspdnce} = Net(\mathcal{X})$;
end

With initial pseudo annotations being very noisy, the model trained at the first rounds should only be able to detect landmarks with low confidence; however, as the model improves, this confidence becomes larger in subsequent rounds. To create the pseudo-ground truth for the next round, a heatmap is created where a small gaussian is placed on each detected landmark. An algorithmic summary of our approach can be found in Algorithm 1.

4.2.3 Training with noisy labels.

A key feature of our approach is to systematically study training under label noise for fine-grained localisation tasks. As identified in [5, 137] deep neural networks can be robust to noisy class labels when trained with the appropriate setting of hyperparameters. To our knowledge, a similar study of training dynamics has not been performed from previous self-training methods that also utilise noisy labels for localisation tasks (like [39, 80] for object segmentation or [193] for saliency detection). We have employed the following strategies:

Explicit Regularisation. Normally, DNNs for landmark localisation are trained

with RMSProp without regularisation [122]. We used regularisation in the form of weight decay equal to $5 \cdot 10^{-5}$. The use of weight decay can limit the speed of memorisation of noise data without significant impact on learning [5].

Larger batch sizes and lower learning rates. We found that large batch sizes (e.g. 128 or more) and a smaller learning rate (10^{-5}) work much better not only for classification tasks [137] but also for fine-grained localisation problems. The rationale behind this is that large batch sizes give better gradient estimates as gradients from random landmarks cancel out. This strategy is in stark contrast to the common tactic of increasing the learning rate linearly with the batch size [88, 55] that has performed well for supervised landmark localisation [122].

Early stopping. Over-parameterized DNNs have enough learning capacity to fully memorise the Superpoint pseudo-labels given sufficient training time. As noted in [5], a DNN would first learn the real underlying patterns and then overfits noise to reduce the training error further. Since our pseudo-labels are similarly a mixture of real landmark locations and noisy background keypoints, our detector would tend to learn the real patterns first (landmark locations). We used early stopping in our training framework by training for only 1500 iterations per iterative round to take advantage of this phenomenon.

Data distillation. Our method also uses data distillation [133], i.e. applies multiple image transformations (crops, rotations and flip) and aggregates the produced heatmaps into a single one before extracting landmarks with NMS. This mechanism can help filter out irrelevant background keypoints as it is unlikely that these can be detected under all possible geometric transformations. We found that data distillation is more effective if applied after the 5-th iteration when the detected landmarks are detected with higher confidence.

Group-Norm vs Batch-Norm. Even though batch-norm is commonly the most effective normalisation method when combined with large batch sizes[68], we find through experimentation that group-norm [177] performs better. A rationale can be that, due

to extreme label noise in every mini-batch, low-quality statistics could be calculated for batch norm despite the large batch size. Hence, a method like group-norm, which does not normalise along the batch dimension, can achieve better performance for training with noisy labels.

Augmentation. We found that strong data augmentation also contributes to regularisation by hindering the memorisation of random labels. Specifically, we applied random flipping, color jittering, scale $(0.7 - 1.3)\times$ and rotation ($-30°$ - $30°$).

4.2.4 Learning correspondence.

Up to this point, the detected landmarks among different images are not in correspondence. Our approach to learning a model (i.e. correspondence) is to specify the desired landmarks we wish to detect by selecting without special effort a single image from the training set to be used as a template and align to it the points detected for the rest of the images in the training set (Fig. 4.2). The alignment can be solved as a set registration problem that allows for a different number of template and detected keypoints. We use the Coherent Point Drift algorithm [121] to align the 2 point clouds and find the correspondence between them by solving the linear sum assignment problem for a cost matrix C where $C[i, j]$ is the Euclidean distance from template point i to keypoint j. Missing landmarks are dealt with by substituting them with the corresponding template points.

Images for which a sufficient number of valid correspondences are found can be used to learn a landmark detection network $\mathbf{\Psi}_\theta : \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{Y}$ now has dimensionality $H_o \times W_o \times K$ with K the number of object landmarks. The new landmark correspondences can still be noisy due to incorrect alignments. Hence, we again use the same iterative training framework for three more rounds with the same hyperparameter setting but 15000 training iterations as we found that a multichannel detector requires a larger training time. In the following Chapters, we will substitute this approach for correspondence recovery through deep clustering.

4.3 Implementation Details

Over each iterative round t, we train a new detector $\mathbf{\Psi}_\theta : \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X} \in \mathcal{R}^{256\times256\times3}$ and $\mathcal{Y} \in \mathcal{R}^{64\times64\times1}$. For $\mathbf{\Psi}_\theta$, we used the HourGlass(HG) [122] architecture with a single network (i.e. no network stacking) and the residual block of [12]. Models are trained with RMSProp and hyperparamters are set as batch size of 128, learning rate of 10^{-5}, weight decay 10^{-5} and 1500 training iterations per iterative round (see Subsection 4.4 for discussion on hyperparameter settings).

For faces, correspondence was recovered using a single template of 13 landmarks (2 on each eye, 2 on the nose and 3 on the mouth), and for human bodies, a template of 7 landmarks (2 on the waist, 1 on each leg and 1 on the head). A template of 5 landmarks (2 around the eyes and 1 on the nose) was used for cat heads. Similarly to [164, 196, 72], our landmark detector is fine-tuned on AFLW by applying self-training for three more rounds.

4.4 Ablation Study

In this section, we evaluate the effect of the proposed hyperparameter settings on the first step of our approach. Several models were trained to detect keypoint locations on the facial dataset CelebA [109] with different settings for regularisation, batch size, learning rate and number of training iterations. To evaluate the first step of our approach, we measure precision and recall in Fig. 4.3 with respect to the 13 ground-truth facial landmark locations our detectors consistency infers (visual example in Fig. 4.6) for facial datasets. Note that correspondence is going to be recovered in the following step, and at this stage, we only evaluate the accuracy [1] of keypoint locations collected through NMS from the single output heatmap.

Round 0 in the depicted figures corresponds to precision and recall values for pseudo-

[1] As a true positive, we consider a keypoint detected within distance $d = 8px$ from one of the 13 supervised groundtruths for an image resolution of 256.

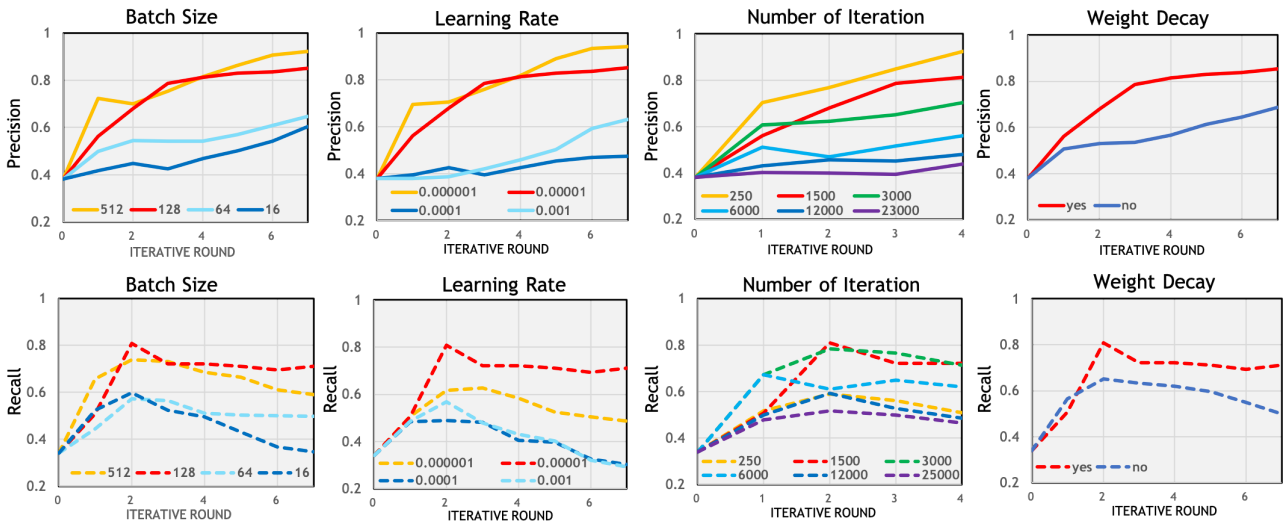


Figure 4.3: Precision curves (top) and recall curves (bottom) per iterative round with respect to 13 groundtruth facial landmark locations. We present results over different hyperparameter settings for learing rate, batch size, number of training iterations and weight decay regularization. The hyper parameters that are not evaluated in each individual experiment are fixed to a good value (learning rate of 10^{-5}, batch size of 128, 1500 training iterations and weight decay of 10^{-5}).

labels detected by the SuperPoint detector. Our assumption that a generic keypoint detector can detect object landmarks with some consistency over images of the same category is reasonably justified with precision and recall values close to 0.4. Moreover, after seven training rounds, both metrics more than double with our iterative training framework. In Fig. 4.3 we see that large learning rates that are commonly used for landmark localisation would result in minimal recall values as the network manages to fit the pseudo-labels very quickly. Similar is the effect of more training iterations per round or very small minibatch sizes that lead to noisy gradient estimates. At the same time, the use of weight decay regularisation also improves performance.

The caveat here is that taking our approach to the extreme with very small learning rates, very large batch sizes or minimum number of training iterations would result in very strong filtering of keypoints with only the most stable being detected by our approach. This is exhibited in Fig. 4.3 with higher precision for a learning rate of 10^{-6} and a batch size of 512 but also small recall as only 8 out of the 13 facial landmarks are recovered. For human faces, these 8 landmarks would be points around the eyes, mouth and nose

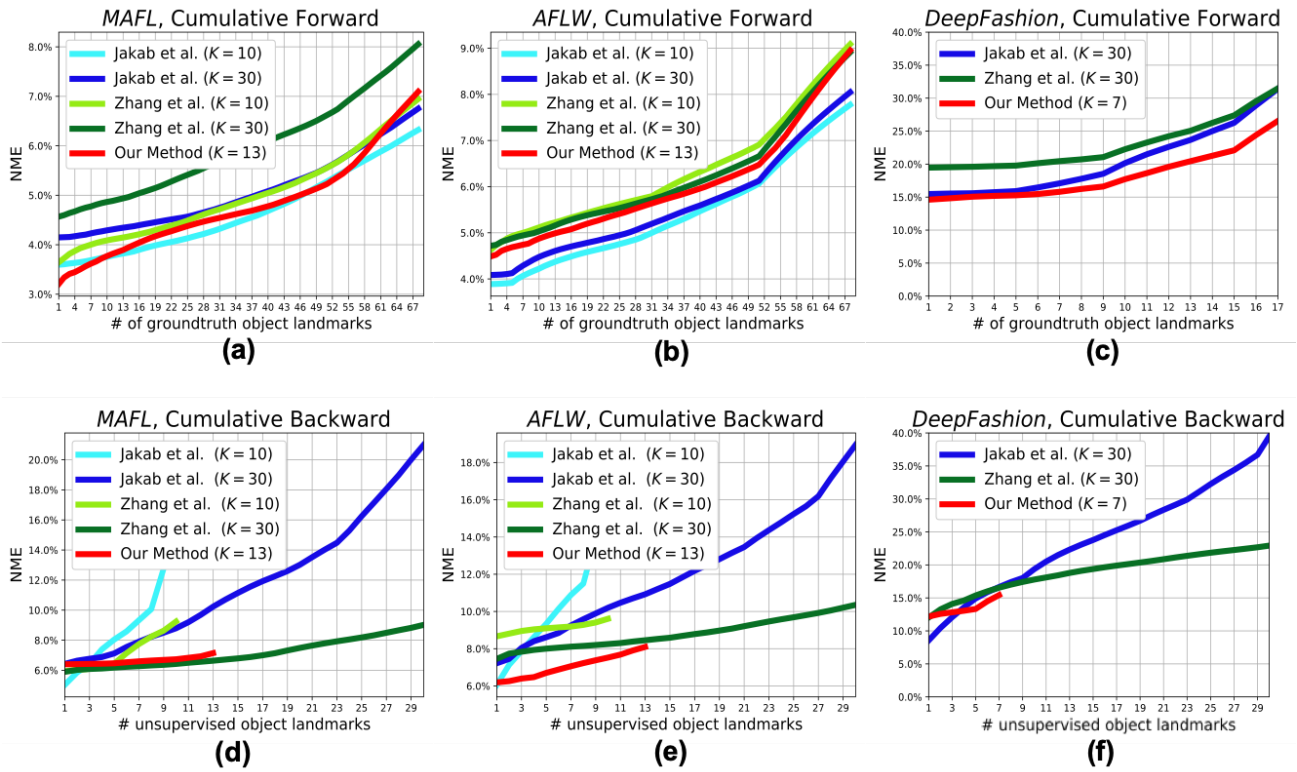


Figure 4.4: Cumulative error curves on various datasets, for both forward and backward NME. For facial datasets we use 68 and for human pose 17 annotated points. Regressor is trained with $N = 200$ training set samples.

since we are not able to detect eyebrows that might be more commonly occluded. We find that the optimal settings are a batch size of 128 with a learning rate of 10^{-5}, with weight decay and 1500 training iterations per iterative round. These settings are used for all our models evaluated in Section 4.5 with no fine-tuning for different datasets.

4.5 Comparison with state-of-the art

The bulk of our results is shown in Fig. 4.4 [2]. Performance for our multichannel landmark detector (Stage 2 of our proposed pipeline) is measured in terms of forward and backward NME (see Subsection 2.6). Our approach offers competitive performance on the forward mapping while mostly surpassing other methods in terms of backward mapping, demonstrating that detected landmarks are highly stable.

[2]Where possible, we used pre-trained models for comparison with other methods. Otherwise, we re-trained these methods using the publicly available code.

Method	MAFL	AFLW
Supervised Methods		
Cascaded CNN[157]	9.73	8.97
TCDCN[202]	7.95	7.65
RAR[158]	-	7.23
MTCNN[201]	5.39	6.90
Unsupervised/ self-supervised Methods		
Thewlis et al. [164] *(K=30)*	7.17	-
Thewlis et al. [164]	5.83	8.80
Shu[149]	5.45	-
Jakab et al.[72] *(K=10)*	3.19	6.86
Zhang et al.[196] *(K=10)*	3.46	7.01
Sanchez [144] *(K=10)*	3.99	6.69
Sahasrabudhe [142]	6.01	-
Ours *(K=10)*	4.44	7.76
Ours *(K=13)*	4.18	7.42

Table 4.1: **Forward-NME** on MAFL and AFLW (normalized by inter-ocular distance) evaluated over 5 groundtruth landmarks. The regression is trained with max N to be consistent with previous work.

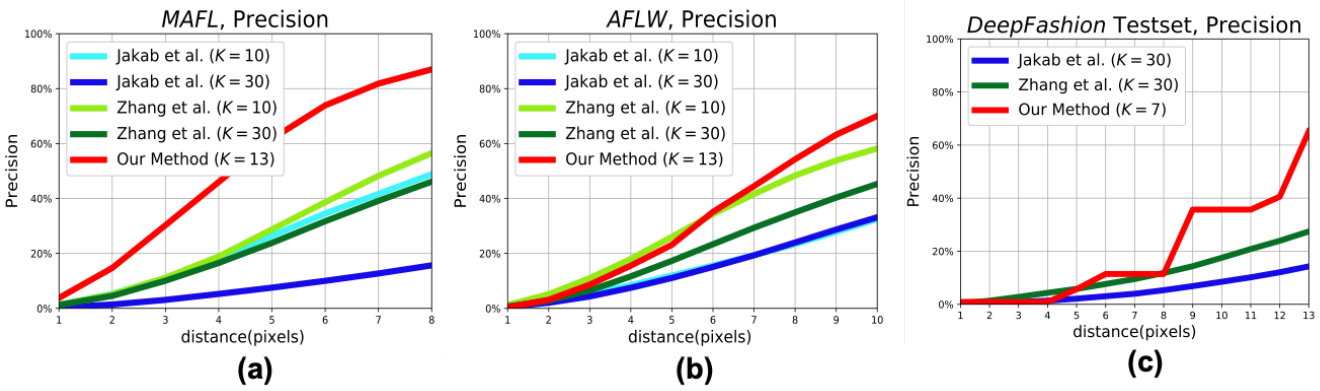


Figure 4.5: Precision measured for unsupervised landmarks with respect to their maximally corresponding supervised ones over a varying distance threshold (image resolution is 256).

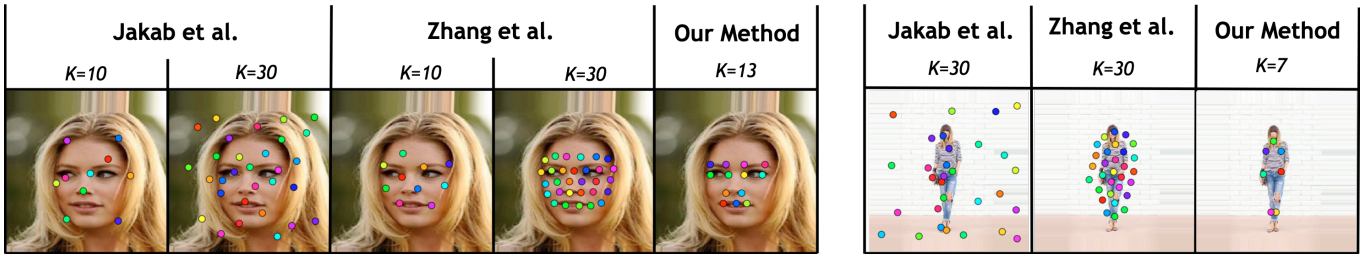


Figure 4.6: Comparison between the landmarks detected by our method and those of [196] and [72] for faces and human poses. Contrary to other approaches, our method is able to discover landmarks with clear semantic meaning.

We compare with approaches that detect both high and low number of landmarks. As mentioned in [164], methods that detect a higher number of landmarks can produce better Forward-NME values as they are more likely to find some unsupervised landmarks that are strongly correlated to the supervised ones. On DeepFashion, we outperform, in terms of the accuracy of the forward mapping, both [72, 196] even though we use a smaller number of unsupervised landmarks.

Note that on MAFL we are even able to match the forward-NME of [72] when we consider up to 52 best-regressed ground-truth landmarks, but then the accuracy of our method degrades faster than other methods. This is because our approach does not detect any landmarks on the border of the face and, as a result, can poorly regress the manually annotated landmarks on the face boundary.

We also report average NME over all 5 always visible landmarks in Table 4.1 for MAFL and AFLW. We measure the performance of our approach using the 10 most stable out of 13 unsupervised landmarks (selected by backward-NME per landmark) for a fair comparison with other methods detecting a smaller number of landmarks. Our approach demonstrates competitive performance for this metric with only a small performance gap (below 1%) with current state-of-the-art methods.

Since our main objective is to discover semantic landmarks, we also assess how well the unsupervised landmarks track semantic landmark locations directly without the need for any mapping function. To this end, we find which ground-truth landmark maximally corresponds to each unsupervised landmark by solving the bipartite linear assignment problem with mean distance as a cost, similarly to [72]. We then measure precision as the percentage of unsupervised landmarks that have been detected within a distance threshold from their maximally corresponding ground-truth on average (Fig. 4.5). We show a precision of over 80% for a distance threshold of 7 pixels on MAFL, showing that our unsupervised landmark detector can be directly used for face alignment with high accuracy without the need for any linear mapping to semantic locations. A visual comparison of the unsupervised landmarks detected by our method compared to other approaches is shown

in Fig. 4.6. Overall our unsupervised landmark detector shows higher precision values in all examined datasets.

4.6 Chapter Conclusions

In this Chapter, we train an object landmark from a generic keypoint without manual supervision. Key features of our method are an iterative training procedure and techniques for training neural nets under extreme label noise. The proposed approach can detect stable landmarks that capture semantic object locations. In the next Chapter we will extend this self-training approach through clustering correspondence to produce a robust framework for unsupervised landmark discovery.

Figure 4.7: Qualitative results for our approach on CelebA.

Figure 4.8: Qualitative results for our approach on AFLW.

Figure 4.9: Qualitative results for our approach on DeepFashion.

Chapter 5

Unsupervised Learning of Object Landmarks via Self-Training Correspondence.

In the previous Chapter, we trained a landmark detector without manual supervision through *self-training* on generic keypoints. Noisy background points were filtered out over iterative self-training, and several techniques for landmark localisation under extreme label noise were explored. Discovered landmarks tracked highly semantic object locations that fire as generic keypoints, and robust performance was demonstrated for various object categories. In this Chapter, we extend the proposed framework through *Clustering Correspondence*. We identify correspondence as a key objective for unsupervised landmark discovery and propose an optimisation scheme which alternates between recovering object landmark correspondence across different images via clustering and learning object landmark representations through the recovered correspondences without labels. Compared to related work, our approach can learn more flexible landmarks in terms of capturing large changes in viewpoint. We show the favourable properties of our method on a variety of challenging datasets.

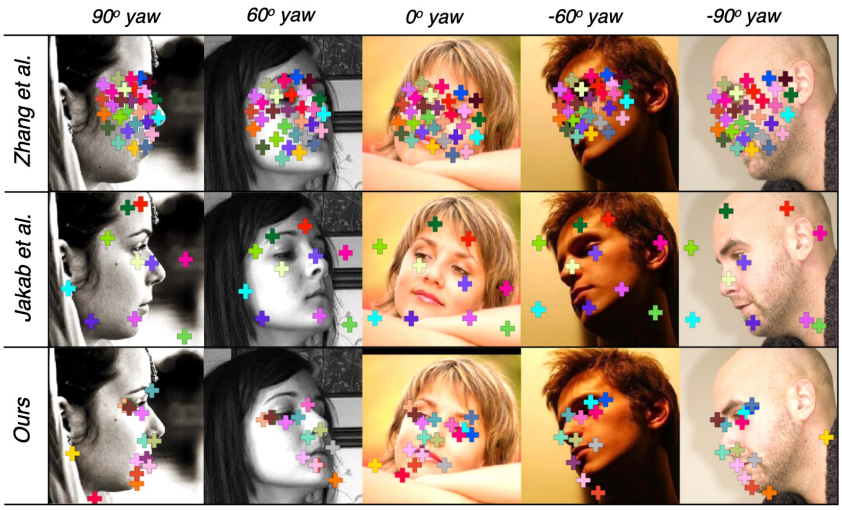


Figure 5.1: Visual comparison between landmarks discovered by our approach and those of [72, 196] on LS3D facial images. Our method is able to both discover highly semantic object landmarks and capture variation in 3D viewpoint across the whole spectrum of facial poses.

The contributions of this Chapter have been published at *NeurIPS*, 2020 in [114]. Code and models are available at `https://github.com/malldimi1/UnsupervisedLandmarks`.

5.1 Motivation

Intuitively, one of the most important properties of a landmark detector is to achieve landmark correspondence across different images of the same category. In the unsupervised case that is the main focus of this thesis, we do not have examples of manually annotated correspondences in our disposal. A landmark detection framework has to discover point correspondences automatically given only raw images from the same object category. As discussed in Subsection 3.1.3 a common approach for correspondence discovery without manual supervision is through synthesised image pairs formed by applying multiple *known transformations* on a single underlying image. We argue that learning point correspondence through synthetic pairs (generated from the same image) results in landmark representations with limited robustness and poor generalisation ability. For example, image-level transformations like colour jittering only weakly simulate the appearance variation of natural objects and viewpoint/shape variations produced from synthetic local wraps (like

the commonly used TPS wrap) cannot capture realistic object deformations.

In the previous Chapter, correspondence between *different object instances* was recovered through alignment with a single point template. Even though this strategy learns landmark correspondence from unpaired data, it has its own limitations. Highly deformable objects or instances captured under large out-of-plane rotations would result in poor alignment with the single point template leading to noisy training correspondences. Moreover, selecting a point template requires some manual user input (we choose a template without special effort). Even though we can achieve better performance through multiple point templates, this approach increases method complexity.

Recently, clustering-based proxy tasks have demonstrated great potential for *self-supervised learning* (see Subsection 3.4). Methods like DeepCluster [19] learn strong image-level representation through iterative self-training where clustering assignments are utilised as pseudo-labels. Motivated by this, we introduce a novel perspective for self-supervised correspondence discovery through deep clustering of landmark representations. To our knowledge, this is the first time that clustering correspondence is applied to the problem of unsupervised landmark discovery.

Our proposed model includes a landmark detector head that extracts landmark coordinates (as in the previous Chapter) and a dense feature extractor that learns local landmark representations. We propose an optimisation scheme which alternates between correspondence recovery via clustering and learning of better correspondence through self-training. Through this clustering approach, we can directly train on unpaired images, in contrast to related methods that can only synthesise correspondences from augmented views of the same underlying image. The proposed formulation results in the discovery of robust landmark representation with stronger invariance to appearance or viewpoint/shape variations.

Compared to previous works, our approach can learn more flexible landmarks in terms of capturing changes in 3D viewpoint. See for example Fig. 5.1. We demonstrate some of the

favourable properties of our method on a variety of difficult datasets, including LS3D [13], BBCPose [24] and Human3.6M [69], notably without utilising temporal information.

5.2 Method

In this section, we will describe the methodology of the proposed approach. We follow the same problem formulation described in the previous Chapter (Subsection 4.2.1) where our initial training set $\mathcal{X}$ is formed as $\{\mathbf{x}_j, \{\mathbf{p}_i^j\}_{i=1}^{N_j}\}$ (where $\mathbf{p}_i^j \in \mathbb{R}^2$ is a keypoint and is N_j the number of detected keypoints in image $\mathbf{x}_j$) and 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 K object landmarks. We will break down our problem into two sub-problems/stages. The first stage aims to establish landmark correspondence, recover missing object landmarks and filter out irrelevant background keypoints. Then, pseudo-labels formed from the first stage will be used to train a strong landmark detector.

5.2.1 Stage 1: Recovering correspondence

We will firstly learn a neural network $\mathbf{\Phi}$ with a shared backbone $\mathbf{\Phi}_b$ and two heads $\mathbf{\Phi}_{h,i}, i = 1, 2$ performing the following tasks:

1. **A detector head** $\mathbf{\Phi}_{h,1} : \mathcal{X} \rightarrow \mathcal{Z}$, where $\mathcal{Z} \in \mathbb{R}^{H_o \times W_o \times 1}$ is the space of single-channel confidence maps, which learns to detect all object landmarks with no order or correspondence, i.e. without distinguishing one from another (hence one output confidence map is used). The main purpose of $\mathbf{\Phi}_{h,1}$ is to recover the originally missed object landmarks. This detector head was also the output of our model in the previous Chapter.

2. **A feature extractor head** $\mathbf{\Phi}_{h,2} : \mathcal{X} \rightarrow \mathcal{D}$, where $\mathcal{D} \in \mathbb{R}^{H_o \times W_o \times d}$ for **recovering correspondence**. For each landmark $\mathbf{p}_i^j$, $\mathbf{\Phi}_{h,2}$ computes a d-dimensional feature

descriptor $\mathbf{f}_i^j$. This descriptor is used to cluster the landmarks into M clusters each meant to represent an object landmark or different views/appearances of the same landmark.

The detector is trained to improve itself without labels through self-training. At every training round t, our method learns $\mathbf{\Phi}_{h,1}$ using the generated pseudo-ground truth landmarks at previous training rounds $t-1$ and $t-2$. The pseudo-ground truth landmarks are simply the model outputs after discarding those that are not close to a cluster centroid (see below). The detector is trained using an MSE loss:

$$L_d = \|H(\mathbf{x}_j) - \Phi_{h,1}(\mathbf{x}_j)\|^2 \tag{5.1}$$

where all landmarks $\{\mathbf{p}_i^j\}_{i=1}^{N_j}$ for image $\mathbf{x}_j$ are represented by a single heatmap H with Gaussians placed at the corresponding landmark locations.

To recover correspondence, we compute the M cluster centroids as well as cluster assignments (using K-means) for the descriptors $\{\mathbf{f}_i^j\}_{i=1}^{N_j}$ collected across all images $\mathbf{x}_j$ in $\mathcal{X}$. We also require that for a specific image no more than one keypoint can be assigned to the same cluster (i.e. object landmark). Hence, a modified K-means problem can be formulated:

$$\min_{C\in\mathbb{R}^{d\times M}} \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_j} \min_{\mathbf{y}_i^j\in\{0,1\}^M} \|\mathbf{f}_i^j - C\mathbf{y}_i^j\|_2^2 \quad \text{s.t.} \quad \mathbf{1}_M^T\mathbf{y}_i^j = 1 \text{ and } \|\sum_j \mathbf{y}_i^j\|_0 = N_j, \tag{5.2}$$

In practice, we solve the above problem (approximately) very fast by running the original K-means to find the cluster centroids and then using the Hungarian algorithm [90] to solve the linear assignment problem between the keypoints for an image and the cluster centroids, ensuring each keypoint is assigned only to a single centroid [1]. Furthermore, keypoints that are not close to any cluster centroid are filtered out. This mechanism enables explicit filtering of noisy keypoints and makes our approach less sensitive to hyperparameter tuning

[1] Clustering followed by the Hungarian algorithm are performed at the end of each training round and are not part of the training the network.

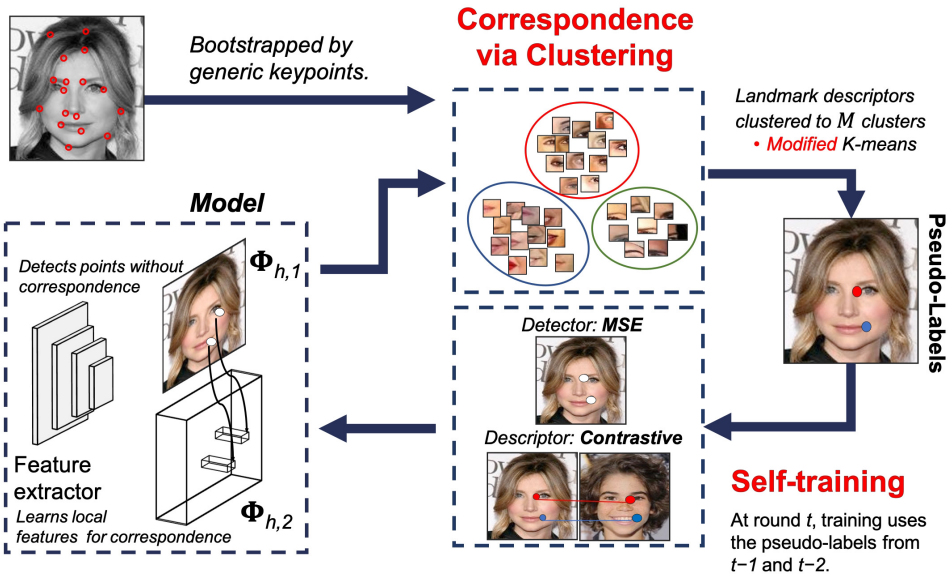


Figure 5.2: Illustration Stage 1 of our proposed landmark detection approach. Our framework is bootstrapped by generic keypoints. During training we alternate between self-training on the pseudolabels of the previous round and forming new pseudo-labels via clustering correspondence.

compared to the work of the previous Chapter.

Under this formulation, K-means would recover M clusters. Note that is crucial to use a sufficiently large M value. This enables our method to *recover several different clusters per landmark* which is necessary as viewpoint changes introduce large appearance changes. This differentiates our approach from prior works, which do not account for large out-of-plane rotations. In a later stage of the algorithm (subsec 5.2.2), we will see how multiple clusters (tracking the same underlying landmark) would be progressively merged, thus enabling the detection of only K underlying object landmarks where $K << M$.

After recovering the keypoint-to-cluster assignments y_i^j, we can directly train $\mathbf{\Phi}_{h,2}$ on unpaired images. We want the features extracted at $\mathbf{p}_i^j$ to be close to those extracted at $\mathbf{p}_{i'}^{j'}$ if and only if $y_i^j = y_{i'}^{j'}$, and far otherwise, i.e. we want $\mathbf{f}_i^j \approx \mathbf{f}_{i'}^{j'} \iff y_i^j = y_{i'}^{j'}$. We formulate this objective in terms of a contrastive loss as:

$$L_c(i, i', j, j') = \mathbf{1}_{[y_i^j = y_{i'}^{j'}]} \|\mathbf{f}_i^j - \mathbf{f}_{i'}^{j'}\|^2 + \mathbf{1}_{[y_i^j \neq y_{i'}^{j'}]} \max(0, m - \|\mathbf{f}_i^j - \mathbf{f}_{i'}^{j'}\|^2), \tag{5.3}$$

where $\mathbf{1}_{[s]}$ is the indicator function, and m is the margin. We form pairs from different

images j, j' as well as by letting j' be a different augmentation of image j. The overall training procedure for $\mathbf{\Phi}_{h,1}$ and $\mathbf{\Phi}_{h,2}$ is based on an alternating optimization and a self-training approach: clustering is performed at the end of each training round, and a new set of pseudo-ground truth locations is added to the training images. An illustration of the first stage of our proposed approach can be seen in Fig. 5.2.

5.2.2 Stage 2: Learning an object landmark detector

Given the pseudo-ground truth landmarks and their cluster assignments $\{\mathbf{x}_j, \{\mathbf{p}_i^j, y_i^j\}_{i=1}^{N_j}\}$ for all images in $\mathcal{X}$ provided by the method of Section 5.2.1, our final goal, in this section, is to train the landmark detector $\mathbf{\Psi}$ (originally defined in Section 4.2.1). To this end, we simply train $\mathbf{\Psi}$ to regress, for each training image $\mathbf{x}_j$, M heatmaps H_m with $m = 1, \ldots, M$ where the m_{th} heatmap is gaussian placed at the pseudo-ground truth landmark location $\mathbf{p}_i^j$ with keypoint-to-cluster assignment y_i^j. For a given image, the model is trained with an MSE loss over all output channels for which there is a landmark-to-cluster assignment for that image:

$$L_d = \sum_m \|H(\mathbf{x}_m) - \Psi(\mathbf{x}_m)\|^2 \tag{5.4}$$

We do not apply the MSE loss for clusters with no landmark assignments. Also, as mentioned in Section 5.2.1, many of the clusters capture the same landmark. Hence, during this step, we also perform *progressive cluster merging*. To this end, we take advantage of the structure of $\mathbf{\Psi}$ (an hourglass network [122]) to set up a simple algorithm which takes into account both appearance and location-related information.

By construction $\mathbf{\Psi}$ can be decomposed into a shared backbone $\mathbf{\Psi}_b$, producing a feature tensor $\mathbf{F} \in \mathbb{R}^{H_o \times W_o \times d}$, and M detectors $\mathbf{w}_m \in \mathbb{R}^{d \times 1}$ (implemented as 1×1 convolutions). If, say, two detectors fire at the same location (k, l) (e.g. for a landmark that might have two close but different descriptors) then, they are both tuned to the same feature

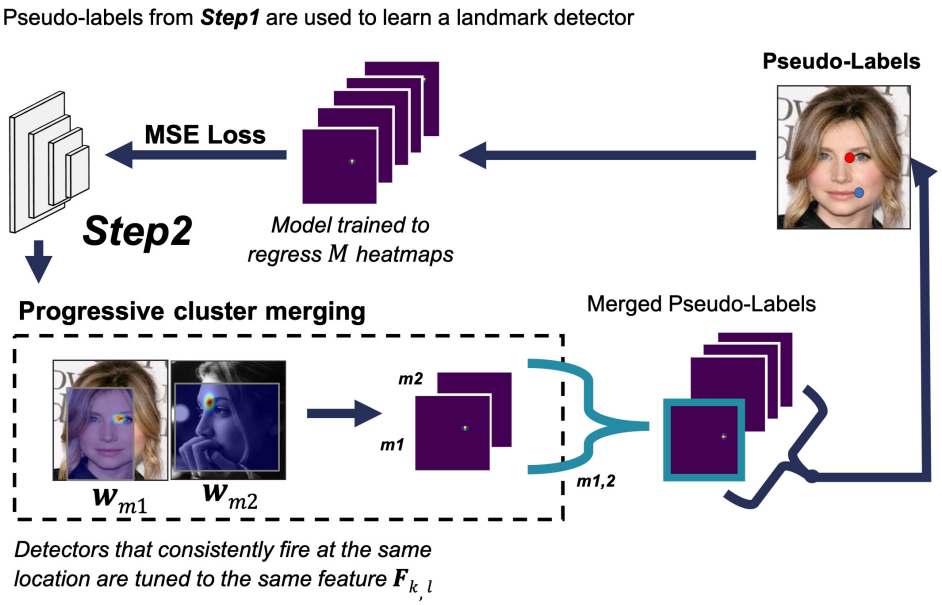


Figure 5.3: Illustration Stage 2 of our proposed landmark detection approach. The landmark detector is trained with standard heatmap regression on the pseudo-labels produced from Stage 1. Progressive cluster merging is applied to group different clusters tracking the same underlying landmark.

$\mathbf{F}_{k,l} \in \mathbb{R}^{d \times 1}$ at that location.

Hence, detectors firing consistently at nearby locations (for the majority of training images) detect features of similar appearance and can be replaced by a single detector. We used this as the criterion for progressively merging different clusters, followed by training. At the end of this step, the number of detectors will be reduced from M to K, where K is the number of landmarks that is automatically discovered by our method. A visual illustration of the second stage of our approach is shown in Fig. 5.3.

5.2.3 Limitations

Thanks to SuperPoint initialisation, our method appears to discover more semantically meaningful landmarks compared to previous methods. However, by no means this guarantees a full semantic meaning for all discovered landmarks. Furthermore, although our approach provides landmarks that can better capture correspondence across large viewpoint changes, in many cases, the discovered landmarks are neither reliably detected

nor able to provide full invariance to large 3D rotations.

Moreover, it is important to remark that the performance of our method greatly depends on the initial keypoint population (see also Section 5.4). On one hand, SuperPoint provides a strong initialisation for our method. On the other, SuperPoint is prone to discovering salient points only, while other semantically meaningful landmarks lying on flat surfaces might be left unpicked. We note though that most object landmark detectors mostly aim to detect salient points (e.g. mouth/eye corners, knee joints). Low texture points are difficult even for the supervised case.

5.3 Implementation Details

5.3.1 Network

We use the Hourglass architecture of [122] with the residual block of [12] 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, with the number of heatmaps being reduced throughout the training as described in Section 6.1.

5.3.2 Training

Keypoints and descriptors are initially populated by SuperPoint [40]. To increase the number of initial landmarks, we applied SuperPoint in 3 scales (1, 1.3 ,1.6). We applied K-means to SuperPoint descriptors to obtain the initial clusters and assignments. For K-means, we used the Faiss library [78]. We also applied an outlier removal step using the same Faiss library (this is not used later in the algorithm). Finally, bounding box information is used to discard the initial keypoints that are detected outside the object of

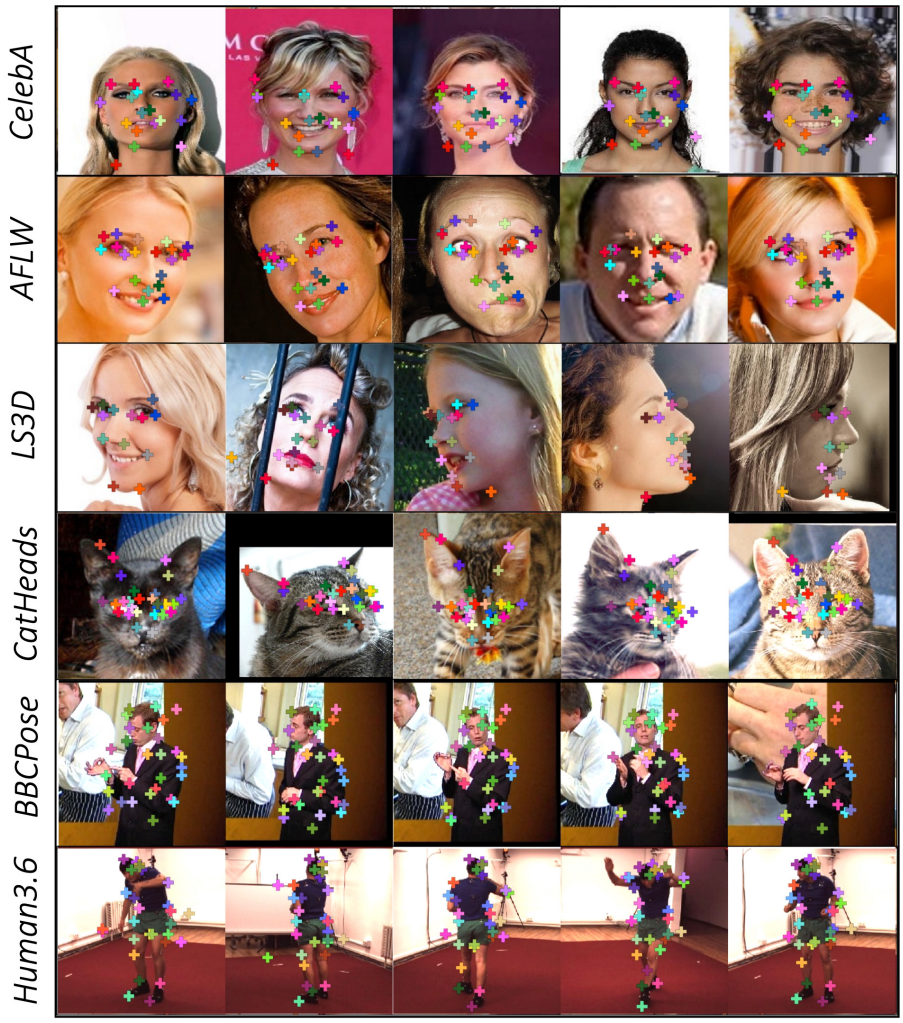


Figure 5.4: Qualitative results of the proposed approach both facial and human pose datasets and cat faces.

interest (this is not used later in the algorithm).

For stage 1, for warm-up, we firstly trained both the detector and the descriptor for 20,000 iterations using as ground-truth the SuperPoint keypoints and their cluster assignments. Then, we trained the model as detailed in Section 5.2.1, applying clustering and updating the pseudo-ground truth every 10,000 iterations. We set $M = 100$ and $M = 250$ clusters for facial and body landmarks, respectively. We found that no more than 300,000 iterations are necessary for the algorithm to converge for all datasets. Due to K-means sensitivity to centroid initialisation, we do not initialise the centroids randomly but use the centroids of the previous training round. During correspondence recovery (and since we always have a larger number of clusters than detected landmarks), to avoid poor

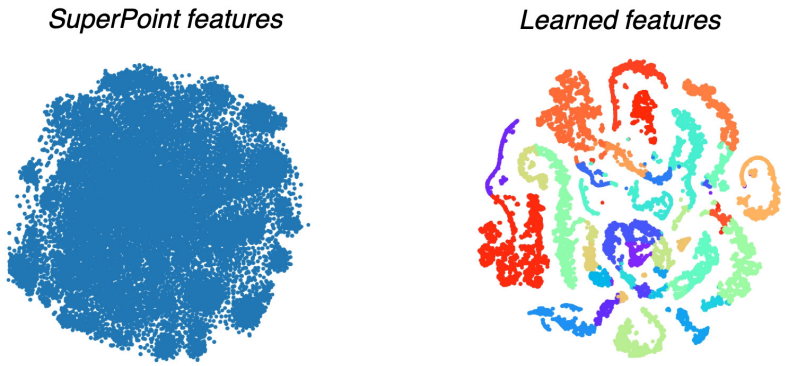


Figure 5.5: Visual comparison between features detected from SuperPoint and our framework with t-SNE [174]

assignments, we discard the assignment of a landmark to a cluster (and the landmark) when its distance to the cluster centroid is much larger compared to the average distance to that centroid.

For stage 2, we initialised the model from the weights of the model of stage 1, except for the weights of the last layer that are trained from scratch. We merged two clusters when 70% of the number of points of the smaller cluster overlap with the points of the biggest cluster (the method does not seem to be so sensitive to this value). Overlap is defined as located within a 1-pixel distance at resolution 64×64. To train the models, we used RMSprop [57], with learning rate equal to $5 \cdot 10^{-4}$, weight decay 10^{-5} and batch-size 16 for stage 1 and 64 for stage 2. All models were implemented in PyTorch [127].

5.4 Ablation Study

5.4.1 Feature representation

We first evaluate the capacity of our method to learn distinctive features corresponding to different landmark locations by computing the t-SNE [174] of the feature representations. Fig. 5.5 shows the t-SNE for the initial SuperPoint features, next to the features returned by our method after self-training. We observe that at the end of training, the descriptors are distinctive of the corresponding classes, making the correspondence recovery effective.

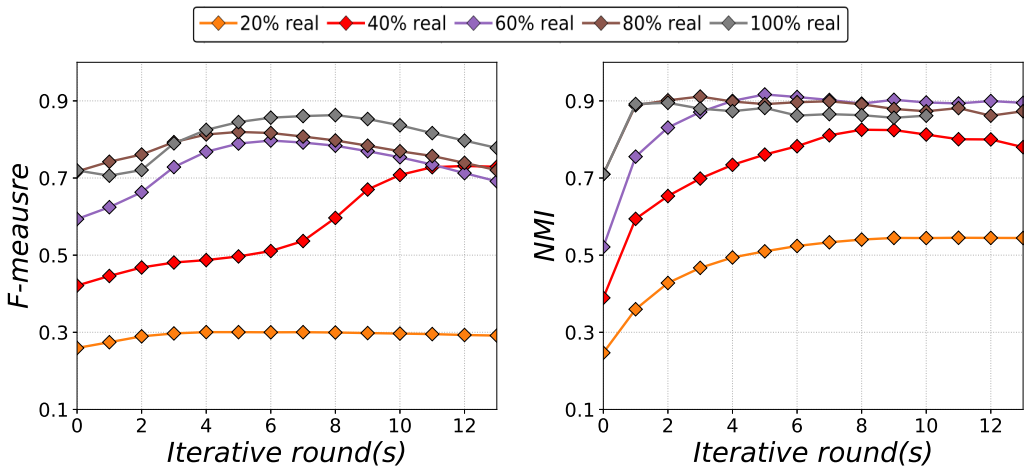


Figure 5.6: Detector and descriptor performance for training with varying mixtures of ground-truth and random keypoints.

5.4.2 Robustness to noise

To study the impact of the quality of the initial keypoints, we replace the generic keypoints (provided by SuperPoint) with a mixture of (1) a varying number of ground-truth points randomly sampled from a set of 15 ground-truth landmarks (3 on the nose and mouth, 2 on each eye and eyebrow and 1 on the jaw) and (2) a set of noise points sampled from the image domain (random 2D locations inside the facial bounding box). The initial number of points per image is randomly chosen to be between 12 and 24, so as to simulate the number of keypoints that are usually returned by SuperPoint. The ground-truth points are also randomly distorted within a 3 pixels radius.

To assess the quality of the pseudo-annotations produced by the first step of our approach, our model's detector and descriptor head are evaluated separately. Since the real ground-truth keypoints are sampled from a specific subset of landmarks, an ideal detector would detect these 15 points in every image and filter out all noise. To evaluate how close our detector is from the ideal one, we measure precision and recall [2] combined with F-measure.

To evaluate the descriptor part of the network, we assess the information shared

[2]As a true positive, we consider a keypoint detected within distance $d = 10px$ to a manually annotated point for an image resolution of 256.

between the clustering assignments produced by our framework and the ground-truth landmark label for each detected keypoint. For a particular keypoint, the landmark label is that of the ground-truth to which it is maximally assigned, and the clustering label is assigned by the clustering of the corresponding descriptors. We measure the Normalized Mutual Information (NMI) between the landmark and clustering assignments. This measure, which is independent of the number of clusters, can quantify the degree to which one assignment is predictable of the other.

The results shown in Fig. 5.6. We observe that when the initial keypoints include more than 20% of ground-truth locations, our method is capable of recovering the right correspondence.

5.4.3 Impact of number of clusters

In addition to the above, we report, in Fig. 5.7, the NMI score for the detector w.r.t. a varying *number of clusters*. Even though keypoints are sampled from 15 groundtruth landmarks, we see that the best performance is attained for $M = 30$. 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.

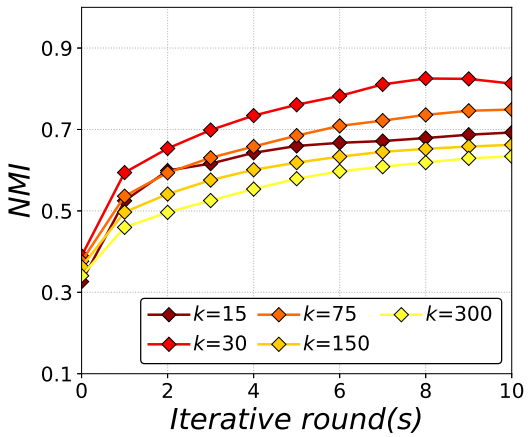


Figure 5.7: Descriptor accuracy by varying the number of clusters. Percentage of real points is 40%.

	CelebA	*AFLW*
R2D2 [135]	4.57	8.87
Superpoint [40]	**4.12**	**7.37**

	CelebA	*AFLW*
Correspondence from Augmented Pairs	0.290	0.352
Clustering for Correspondence	**0.620**	**0.635**

Figure 5.8: *(left)* Forward error for models trained with different keypoint initialization methods. *(right)* NMI for features learned via clustering vs. equivariance.

5.4.4 Keypoint initialization

We evaluate the influence of the initial generic keypoints in our proposed algorithm. To this end, we assess the performance of our method when initialised with the SuperPoint [40], as well as with the recent R2D2 [135], by means of total forward error, evaluated on both CelebA and AFLW. Fig. 5.8 *(left)* shows the results of the forward evaluation of the generic keypoints provided by both methods. We observe that while both yield competitive results, SuperPoint is a better method for initialisation.

5.4.5 Clustering vs. equivariance

We also evaluate the importance of recovering correspondence through clustering of local descriptors. We compare with a model trained without clustering but using image pairs produced by different affine transformations to compute the contrastive loss, i.e. the model was trained with equivariance. In Fig. 5.8 *(right)*, we see that our approach outperforms the previous model by a large margin.

5.5 Comparison with state-of-the art

Herein, we compare our method with [72, 200] trained to discover both 10 and 30 landmarks.

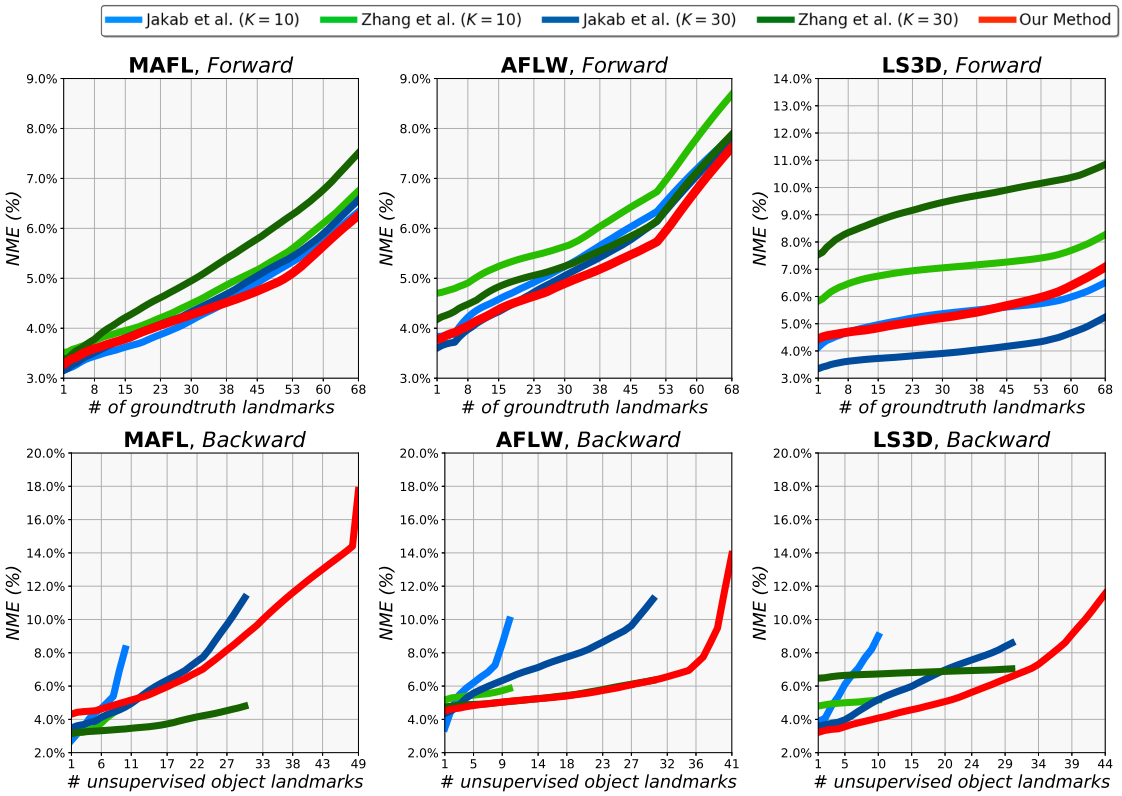


Figure 5.9: Evaluation on facial datasets: Our method discovers 49 landmarks on CelebA, 41 on AFLW and 44 on LS3D. CED curves for forward and backward errors. A set of 300 training images is used to train the regressors. Error is measured w.r.t. the 68-landmark configuration typically used in face alignment.

5.5.1 Evaluation on facial datasets

The bulk of our results on facial images is shown in Fig. 5.9. From our results on all datasets, we can see that our method overall provides the best results in meeting both requirements. Notably, our method delivers state-of-the-art results for the challenging LS3D dataset, which is a dataset with frontal-to-profile pose variations. Table 5.1 shows Forward-NME of our proposed approach compared to various other methods w.r.t 5 facial landmarks where we maintain competitive performance.

Similarly to the previous Chapter, we again find that our approach performs better when evaluation is performed w.r.t the 68-facial landmark configuration compared to 5 facial landmarks. The 5-landmark configuration includes points in uniform areas and not repeatable edges or corners (centre of the eye, centre of the nose), that as discussed in subsection 5.2.3 are not commonly tracked by generic keypoint detectors. Moreover,

Method	MAFL	AFLW
Lorenz [110] *(K=10)*	3.24	-
Shu [149]	5.45	-
Jakab et al.[72] *(K=10)*	3.19	6.86
Zhang et al. [200] *(K=10)*	3.46	7.01
Sanchez [144] *(K=10)*	3.99	6.69
Sahasrabudhe [142]	6.01	-
Ours † *(K=13)*	4.18	7.42
Ours	4.12	7.37

Table 5.1: 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 training images are used to train the regressor and the error is measured w.r.t. to 5 annotated points. †Our approach of the previous Chapter.

qualitative examples on various datasets are shown in Fig. 5.4.

5.5.2 Evaluation on human pose datasets

Performance of our method on the BBCPose and Human3.6M datasets is shown in Fig. 5.10. Our approach demonstrates significantly better accuracy in terms of forward and backward errors for both datasets. 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. This is due to the fact that all methods did not capture the hands of the subject leading to very high error for the corresponding 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

	BBCPose Accuracy (%)					**Human3.6M Accuracy** (%)						
	Head	Shldrs	Elbws	Hands	Avg	Head	Shldrs	Elbws	Waist	Knees	Legs	Avg
Zhang [200]	95.36	32.36	15.02	28.02	42.69	20.90	53.05	50.95	43.70	85.60	1.95	42.69
Jakab [72]	48.29	18.55	19.10	32.11	26.83	00.50	52.15	32.35	26.05	3.70	24.6	23.22
Ours	83.65	62.93	63.13	35.00	61.18	95.1	59.15	58.60	67.59	69.35	73.30	70.35

Table 5.2: Accuracy of raw discovered landmarks that correspond maximally to each ground-truth point measured as %-age of points within $d = 12$px from the ground-truth [72]

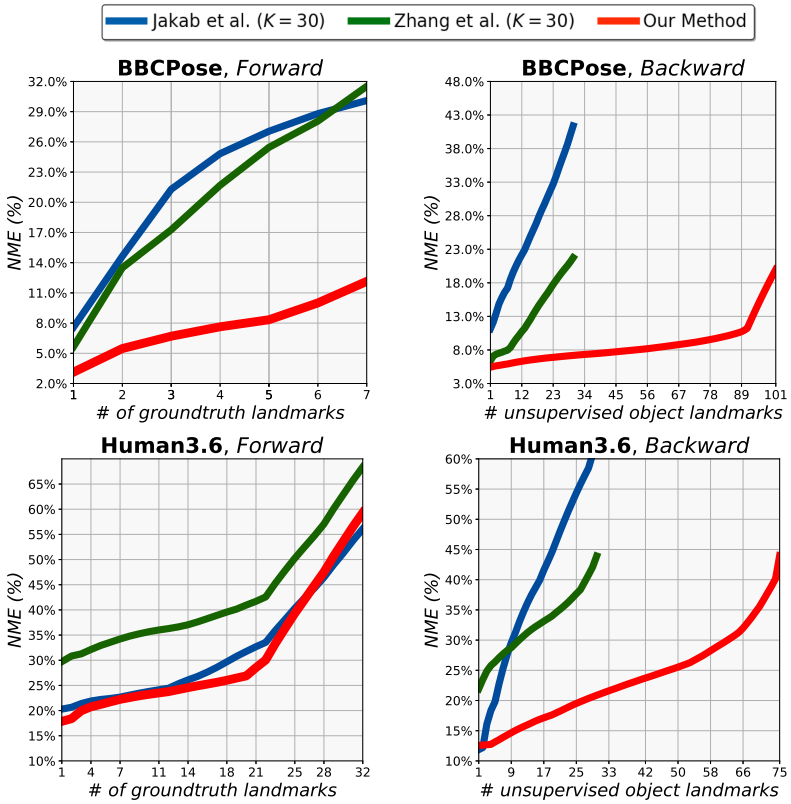


Figure 5.10: Evaluation on human pose datasets: Our method discovers 101 landmarks on BBCPose, and 75 on Human3.6M. CED curves for the forward and backward errors, computed for a regressor trained with 800 samples.

Figure 5.11: Visual demonstration of discovered landmarks (crosses) that maximally correspond to ground-truth keypoints (empty circles).

supervised landmarks. Hence, the forward errors are very high for all methods. To address this, we follow [72] and measure the accuracy of unsupervised landmarks that are found to maximally correspond to the provided ground-truth points (Table 5.2). We can observe that our approach is able to discover clusters that robustly track all parts of the human

body (except the hands for Human3.6M) and show much higher accuracy values compared to the other methods (see Fig. 5.10 for qualitative comparison with other methods).

5.6 Chapter Conclusion

In the Chapter, we presented a novel path for unsupervised discovery of object landmarks based on 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, it has not been previously used for unsupervised object landmark discovery. Compared to recent methods, our approach can learn view-based landmarks that are more flexible in terms of capturing changes in 3D viewpoint, providing superior results on a variety of challenging facial and human pose datasets. The contributions of this Chapter were also presented at the prestigious *NeurIPS2020 conference*.

Figure 5.12: Qualitative results for our approach on AFLW.

Figure 5.13: Qualitative results for our approach on LS3D.

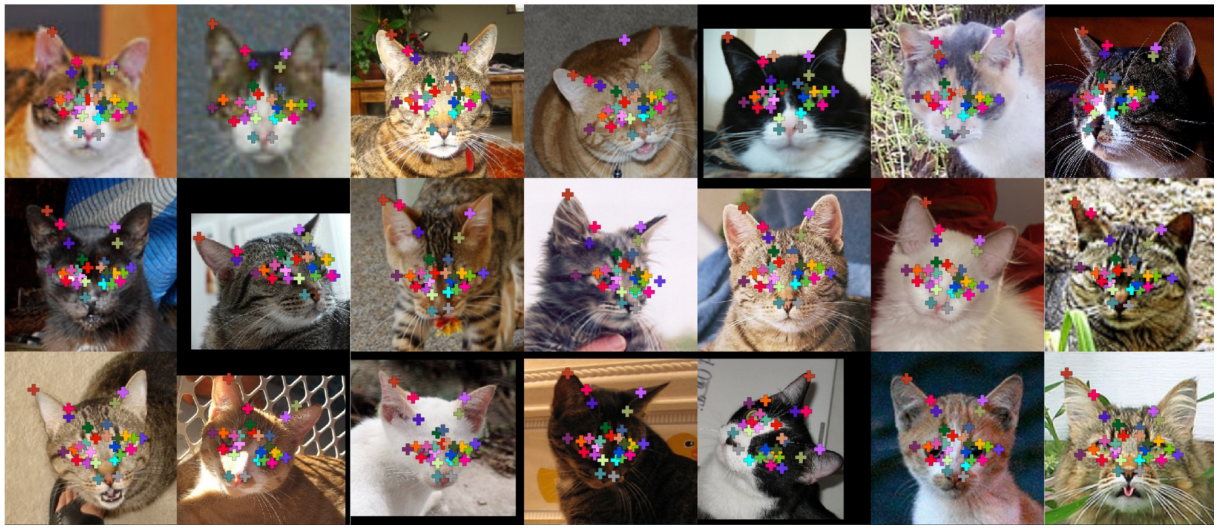


Figure 5.14: Qualitative results for our approach on CatHeads.

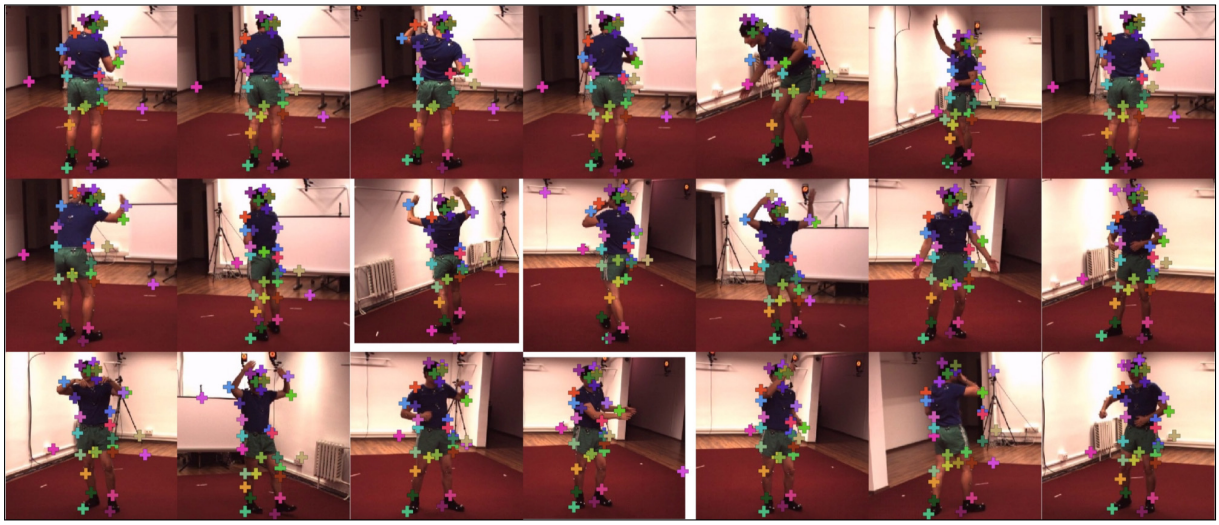


Figure 5.15: Qualitative results for our approach on Human3.6.

Chapter 6

Towards Effective Unsupervised Landmark Discovery.

Previously, we discussed a novel framework for unsupervised landmark discovery. Based on the similarities and differences between keypoints and landmarks, we converted a series of keypoints into semantically coherent landmarks that describe the object parts, filtering and refining during the training process also the corresponding landmark locations. Our approach demonstrated large performance improvement in challenging datasets (LS3D, Human3.6) and impressive robustness to object deformations and large out-of-plane rotations. The proposed framework based on self-training and correspondence via clustering differs significantly from related unsupervised landmark discovery methods that use reconstruction/generation learning objectives. Given that our approach is the first to explore these ideas for unsupervised landmark discovery, we identify that the proposed pipeline can be further refined and potentially simplified. In this Chapter, we rethink various components of our previous work (as presented in the Chapter 5) and discuss how our method can be enhanced through several technical innovations.

This Chapter extends our framework into an effective self-training approach for unsupervised landmark discovery. Our method achieves improved model performance through a simpler training pipeline. We extends our prior work both methodologically

and experimentally. In particular, we introduce the following modifications:

- As presented previously, our method automatically discovered K object landmarks (where K was a result of our algorithm) through the progressive merging of M clusters (see Subsection 5.2.2). This is less flexible compared to other recent unsupervised landmark detectors [164, 196, 72, 144] where K is set apriory (as a hyperparameter). Moreover, we found that the final number of detected landmarks can be sensitive to hyperparameter settings. In this Chapter, we significantly simplify our proposed framework by removing the progressive merging step of our previous work and training our detector to regress K object landmarks directly similarly to prior art.

- We propose a *negative pair selection strategy*, particularly for contrastive learning of object landmark representations. Our approach considers the uniqueness of landmarks, i.e. the fact they can only appear at most once in an image. We experimentally validate that such negative mining enhances learned landmark representations and improves model performance.

- Rather than populating our training set with the descriptors from the pre-trained keypoint detector, we include a warm-up stage to learn the initial features. Such strategy makes the method dependent only on the initial keypoints, regardless of the descriptors.

- We extend our framework to enable flipping augmentation. This augmentation is not utilised for the unsupervised case since landmark correspondence after flipping is unknown. We propose a strategy that recovers flipping correspondence via deep clustering.

Further to the technical innovations described above, this Chapter also contributes a lengthy ablation study that thoroughly analyses our proposed method's different components. We also extend our experimental section and include results for the challenging human pose database PennAction [198] as well as CatFaces [197] and Caltech-UCSD Birds [176].

The contributions of this Chapter have been published at *IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI)*, 2023 in [115]. Code and models can be found at `https://github.com/dimitrismallis/KeypointsToLandmarks`.

6.1 Method

This section will describe our efficient landmark detection framework, focusing on the multiple technical improvements introduced in this Chapter.

6.1.1 Prerequisites

For ease of understanding, we will briefly reiterate the components of our proposed architecture that remain unchanged. As already discussed, Stage 1 of our method learns a neural network $\mathbf{\Phi}$ with a shared backbone $\mathbf{\Phi}_b$ and two heads $\mathbf{\Phi}_{h,i}, i = 1, 2$.

1. **A detector head $\mathbf{\Phi}_1$** will produce, for image $\mathbf{x}_j$, a single-channel spatial confidence map $H_j = \mathbf{\Phi}_1(\mathbf{\Phi}_b(\mathbf{x}_j)) \in \mathbb{R}^{H_o \times W_o \times 1}$ representing the presence/absence of an object landmark at a given location, without regard to any order or correspondence. We use non-maximum suppression (NMS) to extract from H_j the landmark locations $\mathbf{p}_i^j$. The main purpose of $\mathbf{\Phi}_1$ is to recover the originally missed object landmarks.

2. **A feature extractor head $\mathbf{\Phi}_2$** will produce for image $\mathbf{x}_j$ a dense feature map $\mathbf{F}_j = \mathbf{\Phi}_2(\mathbf{\Phi}_b(\mathbf{x}_j)) \in \mathbb{R}^{H_o \times W_o \times d}$ 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.

Recall that after applying $\mathbf{\Phi}$ on the training set, $\mathcal{X}$ becomes $\{\mathbf{x}_j, \{\mathbf{p}_i^j, \mathbf{f}_i^j\}_{i=1}^{N_j}\}$ and correspondence is recovered for each individual keypoint $\mathbf{p}_i^j$ through the clustering assignment of the corresponding feature $\mathbf{f}_i^j$. We refer to this operation as correspondence

recovery, since it allows to identify correspondence of object parts across different images. To assign to each detected keypoint a pseudo-label we follow DeepClustering [19] and perform K-means clustering on the collection of features $\mathbf{f}$.

However, different from [19] where the clusters are used just 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, it is important not to assign two different keypoints on a given image to the same cluster. We also observe that in order to capture landmark descriptors with the same semantic meaning but with possibly different viewpoint-dependent features, it is essential to cluster over a sufficiently large number of clusters M. Thus we formulated the modified K-means problem:

$$\min_{C\in\mathbb{R}^{d\times M}} \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_j} \min_{\mathbf{y}_i^j\in\{0,1\}^M} \|\mathbf{f}_i^j - C\mathbf{y}_i^j\|_2^2 \quad \text{s.t.} \quad \mathbf{1}_M^T\mathbf{y}_i^j = 1 \text{ and } \|\sum_j \mathbf{y}_i^j\|_0 = N_j, \tag{6.1}$$

where M is the number of clusters, $\mathbf{y}_j^i$ is the clustering assignment for landmark $\mathbf{p}_j^i$ and $\boldsymbol{C}$ is the $d \times M$ centroid matrix. Denoting by θ_b, θ_d and θ_f the parameters of $\boldsymbol{\Phi}_b$, $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Phi}_f$, respectively, the full training procedure for Stage 1 is summarised in Algorithm 2.

Algorithm 2: Training of Stage 1

Data: $\mathcal{X}_0 = \{\mathbf{x}_j, \{\mathbf{p}_i^j\}_{i=1}^{N_j}\}$
Compute $\mathbf{y}_i^j$ using Eqn. 6.1
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
    $(\theta_b, \theta_d) \leftarrow (\theta_b, \theta_d) - \nabla_{\theta_b,\theta_d}\mathcal{L}_d$
    $(\theta_b, \theta_f) \leftarrow (\theta_b, \theta_f) - \nabla_{\theta_b,\theta_f}\mathcal{L}_c$
  end
  Update F and p^j using frozen Φ
  Compute $\mathbf{y}_i^j$ using Eqn. 6.1
  Update $\mathcal{X}_t = \{\mathbf{x}_j, \{\mathbf{p}_i^j, \mathbf{y}_i^j\}_{i=1}^{N_j}\}$
end

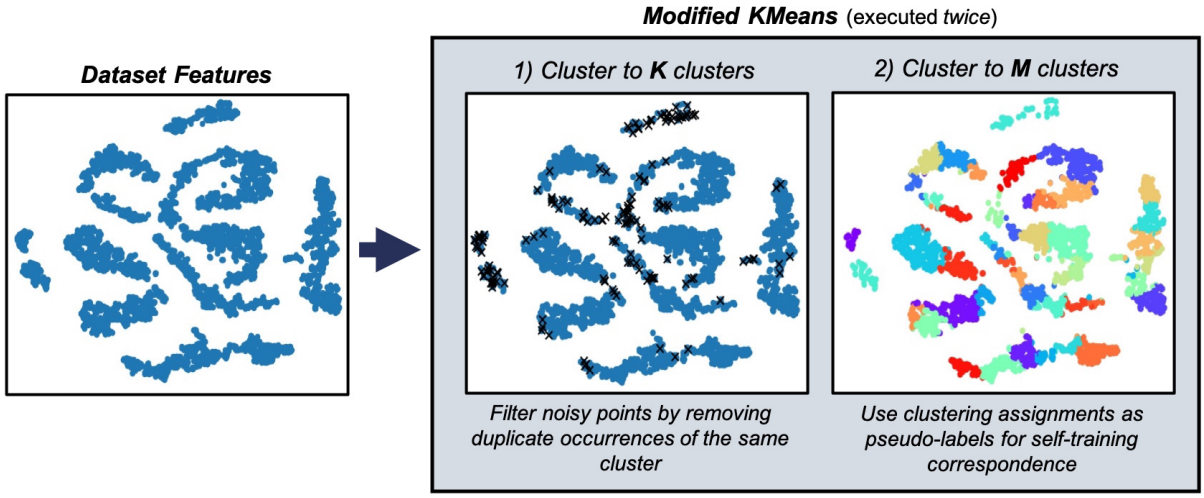


Figure 6.1: Recovering correspondence and ensuring the detection of at most K keypoints per image (illustrated through t-SNE[174] visualisation of algorithm steps). Modified K-means is executed *twice*. The *first time* we cluster to K clusters to filter out duplicate cluster occurrences in a single image (we mark with ✕'s the keypoints that get filtered out). The *second time* we cluster the reduced set to M clusters to enable our method to recover multiple clusters per object landmark.

6.1.2 A simpler implementation of Modified K-means

The modified Kmeans problem was approximately solved by running the original K-means to find the cluster centroids and then using the Hungarian algorithm [90] to computer the linear assignments between the keypoints for a particular image and the cluster centroids. This ensured the constraint $\| \sum_j \mathbf{y}_i^j \|_0 = N_j$ is satisfied with each keypoint optimally assigned to the most representative centroid.

In this Chapter, we identify that given the self-training nature of our proposed method, multiple detected keypoints can be noisy, especially in early training. A straightforward filtering approach is to simply drop duplicate occurrences of the same cluster for a single image (and not assign them to the next more representative centroid as previously). 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. To that end, we satisfy the $\| \sum_j \mathbf{y}_i^j \|_0 = N_j$ constraint and find $\{\mathbf{y}_i^j\}$, $\{\mathbf{f}_i^j\}$ by simply removing duplicate occurrences of the same cluster k on a single image (only keep the feature $\mathbf{f}_i^j$ that is further closest to the centroid).

6.1.3 Detection of at most K keypoints per image

An intuitive way of constraining the detector to at most K keypoints per image would be to use $M = K$ clusters directly. Since the modified K-means drops duplicate occurrences of the same cluster, this filtering mechanism would progressively constrain the detector head into discovering K keypoints per image. As we have discussed though, it is crucial to use $M \gg K$. The resulting over-segmentation of the feature space enabled our method to recover several different clusters per landmark, which is necessary as viewpoint changes introduce large appearance changes.

We can enable our method to both detect K landmarks per image and over segment the feature space to M clusters with $M \gg K$ by simply executing modified K-means *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 fewer 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. These M clustering assignments will be used for correspondence learning during the following iterative round. The whole process is illustrated in Fig. 6.1. Note that even though clustering is performed twice, using an accelerated similarity search method [78] this step can be executed very fast.

6.1.4 Negative pair Selection

Recall that following correspondence recovery at round t, $\mathcal{X}_t$ becomes $\{\mathbf{x}_j, \{\mathbf{p}_i^j, \mathbf{y}_i^j\}_{i=1}^{N_j}\}$, These clustering correspondences are going to be used to train $\mathbf{\Phi}_{h,2}$ on the subsequent training iteration $t+1$ using a contrastive loss. 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. Landmarks assigned to the same cluster constitute natural positive pairs and as a result we want features $\mathbf{f}_i^j$ extracted at $\mathbf{p}_i^j$ to be

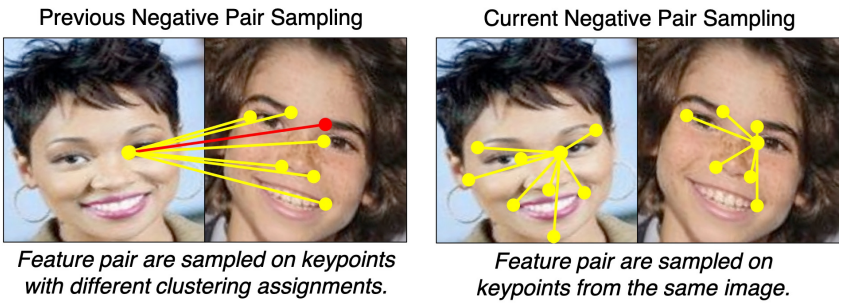


Figure 6.2: Comparison of negative selection strategies. In the previous Chapter, negative pairs were sampled on keypoint locations with different clustering assignments. Since multiple clusters can track the same landmark, this can lead to inaccurate negative pairs *(red line)*. Sampling negatives from the same image guarantees accurate pairs, given that, by definition, each landmark can only appear once per image.

close to $\mathbf{f}_{i'}^{j'}$ extracted at $\mathbf{p}_{i'}^{j'}$ if and only if $y_i^j = y_{i'}^{j'}$.

On the other side, negative pairs can be chosen in many different ways. In our previous work, negative pairs where selected as features $\mathbf{f}_i^j$,$\mathbf{f}_{i'}^{j'}$ with $y_i^j \neq y_{i'}^{j'}$ (assigned to different clusters). This led to a contrastive loss formulation:

$$L_c(i, i', j, j') = \mathbf{1}_{[y_i^j = y_{i'}^{j'}]} \|\mathbf{f}_i^j - \mathbf{f}_{i'}^{j'}\|^2 + \mathbf{1}_{[y_i^j \neq y_{i'}^{j'}]} \max(0, m - \|\mathbf{f}_i^j - \mathbf{f}_{i'}^{j'}\|^2), \tag{6.2}$$

where $\mathbf{1}_{[s]}$ is the indicator function, and m is the margin (as also discussed in Subsection 5.2.1).

Here we improve our landmark representation performance through a novel strategy for negative pairs selection based on the structure of the landmark detection task. Our approach is to select negative pairs *from the same image only*. Given the over-segmentation of the underlying landmarks to M clusters where $M \gg K$, a single landmark is guaranteed to be captured by multiple clusters tracking various appearance/pose variations. The result is that when $y_i^j \neq y_{i'}^{j'}$ features $\mathbf{f}_i^j$, $\mathbf{f}_{i'}^{j'}$ can be either of different underlying landmarks or of *a different cluster tracking the same underlying landmark*. Due to this uncertainty, features from different images are not good candidates for negative pairs.

On the other hand, given the structure of the landmark detection task, each object

landmark can only appear once per image. Features $\mathbf{f}_i^j$, $\mathbf{f}_{i'}^j$ are always guaranteed to be from different underlying landmarks and can be informative negative pairs. This can be formulated as:

$$L_c(i, i', j, j') = \mathbf{1}_{[y_i^j = y_{i'}^{j'}]} \|\mathbf{f}_i^j - \mathbf{f}_{i'}^{j'}\|^2 + \mathbf{1}_{[y_i^j \neq y_{i'}^j]} \max(0, m - \|\mathbf{f}_i^j - \mathbf{f}_{i'}^j\|^2), \tag{6.3}$$

where the only difference compared to equation 6.2 are the indices for the negative samples. Similarly to the previous Chapter, we form pairs from different images j, j' as well as by letting j' be a different augmentation of image j. In practice, to increase the number of negative pairs, we also include negative pairs with features at random background locations. An illustration of our negative pair mining strategy is shown in Fig. 6.2.

6.1.5 Warm up

Initially, at round $t = 0$ our 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 the contrastive loss as described in the previous section. In our previous work, point correspondences were initially recovered using features produced by a generic keypoint descriptor like SuperPoint [40]. This work includes a bootstrapping stage where we train the feature extractor form pairs of images j, j' where j' be a different augmentation of image j. We form known point correspondences between the versions of the same image through synthetic augmentations that can be used as initial positive pairs. Essentially, we first warm-up our method using equivariance training and then further improve our feature extractor through learning from unpaired data (with clustering correspondence). This bootstrapping stage limits our method's initialisation requirements to just generic keypoint locations without the need for the corresponding descriptors.

6.1.6 Learning K clusters

Given the pseudo-ground truth landmarks and their cluster assignments for all images in $\mathcal{X}$, our final goal in this section is to train the landmark detector $\mathbf{\Psi}$ (initially defined in Section 4.2.1). In our previous work [114], cluster assignments where w.r.t M clusters (with $M >> K$) and a progressive merging step was required to combine clusters tracking the same underlying landmark. In this work, even though not explicitly enforced, we find that through self-training learned features automatically form K well-separated clusters, thus lifting the need for a progressive cluster merging step. This is due to *(1)* constraining the detector on at most K keypoints per image and *(2)* the use of our negative pair sampling strategy. Since at most K keypoints are detected per image that are all encouraged to have well-separated features (only features from the same image are used as negative pairs), K clusters emerge automatically over the whole dataset. An illustration of the complete framework for Stage 1 of our proposed approach can be seenn in Fig. 6.3. During training, features are clustered to M clusters with $M >> K$, whereas after self-training is completed, the learned representations are well separated to K clusters.

6.1.7 Training a landmark detector

For Stage 2 of our approach, we simply use the pseudo-groundtruth of Stage 1. To finally populate our training set with K clusters only, we perform a last K-means clustering with K clusters. We train $\mathbf{\Psi}$ to regress, for each training image $\mathbf{x}_j$, 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. Since keypoints are directly clustered to K clusters, progressive cluster merging is not required. For a given image, the model is trained with an MSE loss over all output channels for which there is landmark-to-cluster assignment for that image: $L_d = \sum_k \|H(\mathbf{x}_k) - \Psi(\mathbf{x}_k)\|^2$. We do not apply the MSE loss for clusters with no landmark assignments.

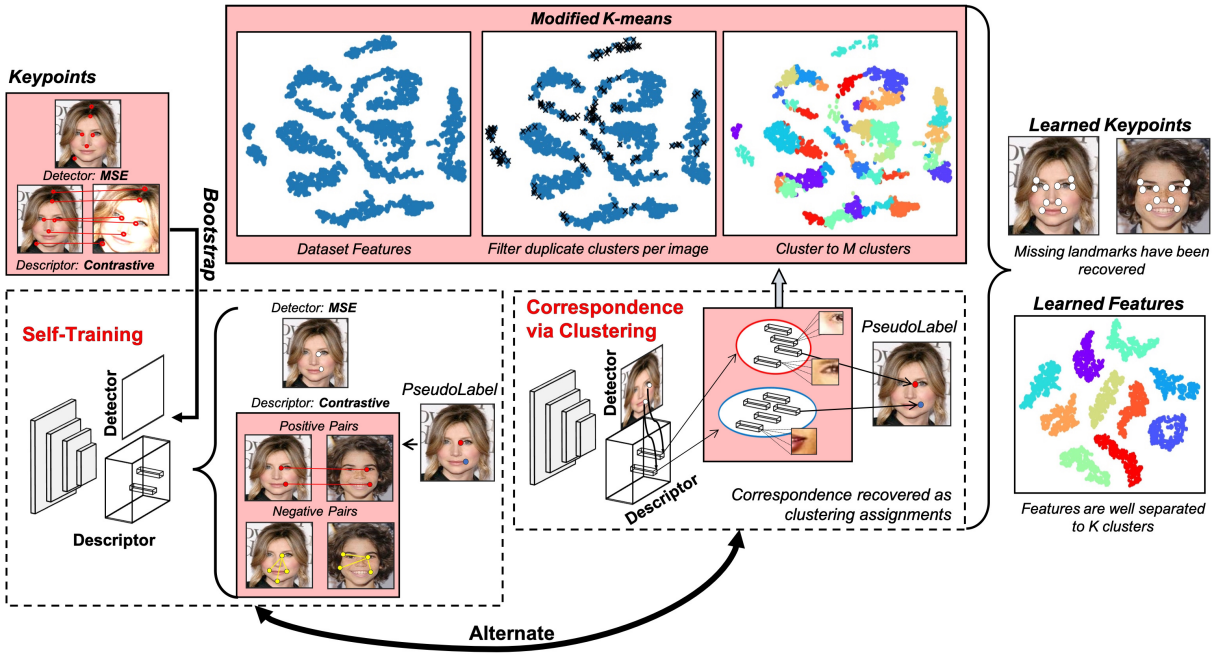


Figure 6.3: **Stage 1** of our Efficient Landmark Detection approach. Novel framework components are highlighted. We improve upon our previous work by *(1)* bootstrapping through equivariant training (Subsection 6.1.5), *(2)* constraining our model to detect at most K keypoints per image (Subsection 6.1.3), *(3)* a novel negative selection strategy (Subsection 6.1.4). Our framework progressively learns K well separated clusters that can be used to train a full landmark detector *without the need for progressive cluster merging*.

6.1.8 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. We propose to recover symmetric landmark correspondence using deep clustering. At the correspondence recovery step (subsection 6.1.1), for each detected keypoint, features are sampled on both the original image and the flipped version of an image (on the corresponding keypoint location). We treat these features independently and produce two clustering assignments for each keypoint (one for the original and one for the flipped image). During training of Stage1, the clustering assignments of the flipped feature is 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.

6.2 Implementation Details

6.2.1 Training

Keypoints and descriptors are initially populated by SuperPoint [40]. For K-means, we used the Faiss library [78]. We also applied an outlier removal step using the same Faiss library (this is not used later in the algorithm). We perform warm-up for 30,000 iterations as described in 6.1.5. 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 examined datasets. Contrary to our previous work, we do not use a larger number M for human pose datasets. We found that no more than 200,000 iterations are necessary for the algorithm to converge for all datasets. For Stage 2, we initialised the model from the weights of the model of Stage 1, except for the weights of the last layer that are trained from scratch. To train the models, we used RMSprop [57], with learning rate equal to $2 \cdot 10^{-4}$, weight decay 10^{-5} and batch-size 16. All models were implemented in PyTorch [127].

Similarly to other recent methods [72, 200] we also train models with additional temporal supervision (optical flow for short-term self-supervision as in [200]) for experiments on video datasets. Note that our approach achieves good performance without temporal supervision, and optical flow is used only when explicitly stated.

6.2.2 Evaluation

Our landmark detector (after Stage 2) directly regress K object landmarks, and evaluation is performed following the standard metrics described in Subsection 2.6. Further to present

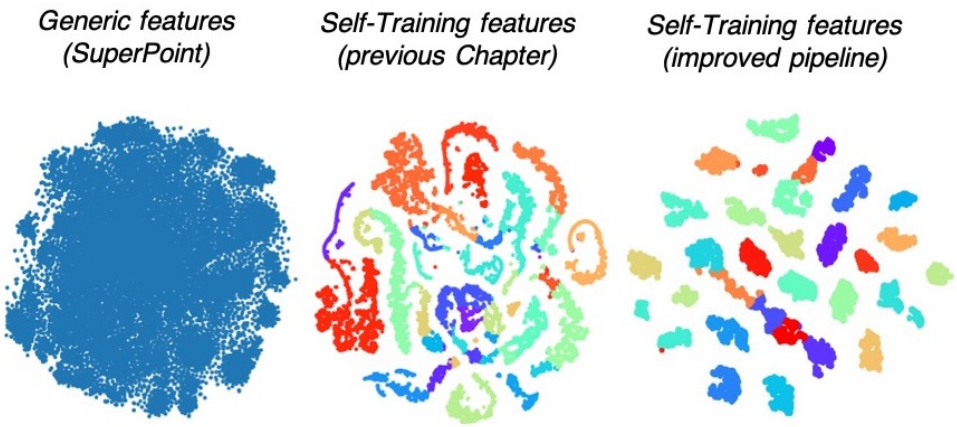


Figure 6.4: T-SNE[174] visualisation of local features. Comparison with features produced by SuperPoint[40] as well as the pipeline discussed in the previous Chapter.

evaluation for the object landmark detector learned from the Stage2 of our approach (Subsection 6.1.6), we also conduct experiments where keypoints and correspondences learned from Stage1 are evaluated directly (Section 6.3 *Ablations*) without learning a full landmark detector. Since each cluster does not appear in every image under our first stage formulation, we have to train a regressor to predict ground-truth annotations from a varying number of visible landmarks. We opt for completing the missing values in the training set with the Singular Value Thresholding method for Matrix Completion [15] (leaving the detected points unchanged). For a given test image, we fill the missing values by the average landmark for that position, calculated from the training partition.

6.3 Ablation Study

This Subsection included an extended ablation study on our efficient landmark detection framework.

6.3.1 Feature representation:

As in the previous Chapter (Subsection 5.4.1), we start by qualitatively evaluating the capacity of our method to learn distinctive features by computing the t-SNE [174] of the

feature representations. Here we compare in Fig. 6.4 with the t-SNE for generic SuperPoint features, as well as features returned by the framework of the previous Chapter after self-training. We observe that our efficient landmark detection framework learns more distinctive descriptors. This allows the removal of the progressive merging step since features can be directly clustered to K clusters.

6.3.2 Robustness to noise

Even though generic keypoints capture several object landmark locations on images of a specific object category, they also include a large number of noisy non-corresponding points. The proposed landmark detection framework of the previous Chapter has already demonstrated strong robustness to initialisation noise (Subsection 5.4.2). To evaluate the effect of this noise on our effective landmark detection approach, we repeat the synthetic experiment of Subsection 5.4.2 where instead of keypoints, we bootstrap our method by a mixture of actual ground-truth landmark locations along with noisy points randomly sampled from the image domain. We follow the same setup (Subsection 5.4.2) where keypoint initialisation for each image is a set of 15 points that are sampled either from the ground-truth locations of 15 facial landmarks (eyes, eyebrows, nose, mouth, chin) or as a random image location. 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.

In Subsection 5.4.2, the detector and feature extractor were evaluated separately by F-measure and NMI where NMI allowed for an evaluation that was independent to the number of clusters (recall that in our previous work, keypoints were separated into M clusters that get progressively merged over Stage 2). Since we can now constrain our model in detecting $K = 15$ (equal to the number of ground-truth landmarks used in this synthetic experiment), we opt for a simpler evaluation process where performance is measured in terms of total forward error. Results of this experiment are shown in Fig. 6.5. Similarly to 5.4.2, we find that even with as much as only 20% of real object landmarks in the

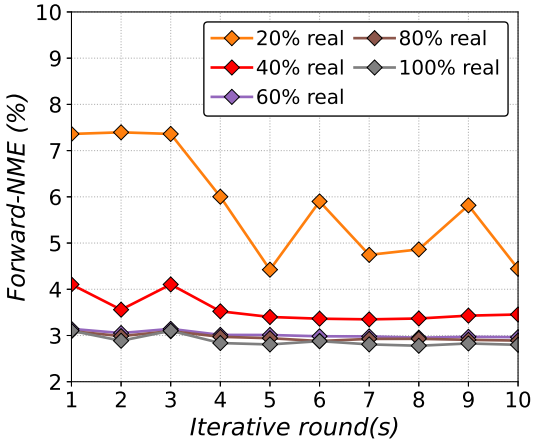


Figure 6.5: Forward-NME (shown for the first 10 iterative rounds) of training the first stage of our method with varying ratios of real and random points. Experiment is performed on CelebA [109]. Real points are sampled from 15 facial landmarks and further perturbed spatially by a small offset sampled from $[-3px, +3px]$.

# 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 6.1: Evaluation of landmarks learned from the first stage of our approach on LS3D [13] 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 variations of the same landmark to be captured by several clusters

keypoint initialisation, our method is still able to recover to some extent. Increases in the percentage of real points over 40% only result in minor performance gains.

6.3.3 Impact of number of clusters

We report, in Table. 6.1, performance for a detector trained with varying *number of clusters*. As in the previous experiment, we again report forward/backward error instead of NMI and F-measure. We verify the findings of the previous Chapter since, even though our model is trained to detect $K = 30$ object landmarks, we see 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 6.1).

Here we also observe that 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 the ability of our method to learn representations that are robust to viewpoint or appearance variations, and more diverse landmarks get filtered out (leading to an increase in Forward-NME). Note that, in the extreme case where M is equal to the number of detected keypoints (each cluster contains a only single feature), our method essentially equates to equivariance training. Even though we could potentially optimise the cluster number of different datasets, we opt for consistently using 100 clusters for all experiments.

6.3.4 Keypoint initialisation

In this work, we use SuperPoint [40] to provide keypoint initialisation for our method, but other keypoint detectors can also be used. In the previous Chapter (Subsection 5.4.4) we contrasted SuperPoint [40] to the recent R2D2 [135] method. Here we expand our analysis to also include as popular SIFT [112] and ORB [139] detectors. Note that SIFT and ORB tend to detect large numbers of spatially clustered points, thus providing a suboptimal initialisation of landmark detection. Moreover, since we perform *modified* K-means across keypoint features collected over the whole dataset (see Section 6.1.1), a large number of spatially clustered keypoints per image can increase computational requirements. Even though similarity search methods like Faiss [78] can scale to very large datasets (through lossy compression based on product quantizers), here we opt for simply combining SIFT and ORB with the adaptive non-maximal suppression method of [8], thus ensuring a homogeneous spatial distribution and constraining the number of detecting keypoints.

Keypoint Detector	Forward	Backward
SIFT [112] + *ANMS* [8]	4.07	7.79
ORB [139] + *ANMS* [8]	3.85	7.70
R2D2 [135]	3.71	7.97
SuperPoint [40]	**3.25**	**6.65**

Table 6.2: Evaluation of landmarks learned on the first stage of our framework on CelebA under different keypoint initialisation methods. Models are trained to for $K = 30$.

Table. 6.2 shows the results of the forward evaluation of the generic keypoints provided by the examined methods. While all methods yield competitive results, SuperPoint is a better method for initialisation. Note that handcrafted methods were not included in the similar analysis of the previous Chapter (Subsection 5.4.4) as it was observed that they did not provide a sufficiently strong initialisation. More specifically, generic features provided by handcrafted methods could not capture any invariance to appearance or viewpoint variations of the same landmark and resulted in poor initial correspondence recovery. Due to the warm-up stage introduced in this Chapter, we no longer require generic features to bootstrap our method, and handcrafted keypoint detectors can also used to provide initialisation that leads to competitive performance (even thought SuperPoint remains the strongest initialisation method).

6.3.5 Negative-Pair Selection

We conduct an ablation study on the proposed negative pair selection strategy (referred to as *same image only*). We compared to the approach described in the previous Chapter (referred to as *different cluster*) where negative pairs were selected as keypoints with different clustering assignments. Results in Table 6.3 are given with both correspondence recovery through *clustering* and *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), but for training through the contrastive loss we only use augmented views of the same image.

We observe that our improved negative pair selection strategy is the best performing

#	**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 6.3: Ablation study of the proposed negative pair selection strategy (compared to the strategy of [114]), combined with either clustering or equivariance training. Experiment performed in the challenging LS3D [13] 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 6.4: 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.

method when correspondence is recovered via clustering (*line 4*). The *different cluster* strategy separates features into M clusters (*line 1*) and results in poor performance when it is not combined with an additional merging step (as in the previous Chapter). 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. 6.2) do not emerge. As a result negative pairs from *different cluster* are more informative and result in better performance (experiment in *line 3* performs better than experiment in *line 2*).

6.3.6 Stage 1 vs 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

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 6.5: 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.

	LS3D-Test				*CelebA-Test*
TrainSet	$[0°, 30°]$	$[30°, 60°]$	$[60°, 90°]$	total	total
CelebA	5.86	6.21	9.20	6.95	3.25
LS3D	4.42	5.07	6.51	5.26	3.33

Table 6.6: Cross-Dataset evaluation. We report the forward-NME on the test partitions of CelebA and LS3D. For LS3D-Test (LS3D-Balanced), error is shown across poses (measured in buckets of different yaw angles).

number of detected points per image in 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 6.4 we compare the 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, which results in lower error values. Performance increase is most notable on LS3D, where occlusion is extended due to large jaw angles.

6.3.7 Generalisation

To further analyse the robustness of our proposed approach to in-plane rotations as well as to domain shift, we conduct a cross-dataset evaluation (Table 6.6). In particular, for $K = 30$, we evaluate models trained on CelebA and LS3D on the corresponding test partitions (LS3D-Balanced and MAFL). In Table 6.6, we break down the results for LS3D-Balanced in different yaw angle ranges (to allow for a better comparison between the

same-dataset results shown in Table 6.4 and those given by this cross-dataset evaluation). We can observe that due to the lack of in-plane rotations on CelebA, the model tends to produce high error values for larger poses in LS3D-Test. On the contrary, the model trained on LS3D can maintain its robustness on the CelebA-Test partition, given that it is composed of mostly frontal faces. We can also observe that the improvement of the LS3D model compared to that trained on CelebA, is quite significant on the most difficult partition of the LS3D-Test, illustrating the need of having a diverse set of images describing the geometry of the target object.

6.3.8 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 6.5. We observe that both flipping as a training augmentation and flipping at test time result in consistent performance improvement.

6.4 Comparison with state-of-the art

This Section presents the experiments carried out to validate the proposed approach. Some qualitative results on various datasets are shown in Fig. 6.6.

6.4.1 Evaluation on facial datasets

We present results for our method trained for both 10 and 30 landmarks. Results in Table. 6.7 show the commonly reported forward error w.r.t 5 ground-truth facial landmarks. Even though methods that detect a higher number of landmarks can result in better Forward-NME values (as also mentioned in [164]), our effective landmark detection framework with $K = 10$ improves performance on this evaluation compared to our previous

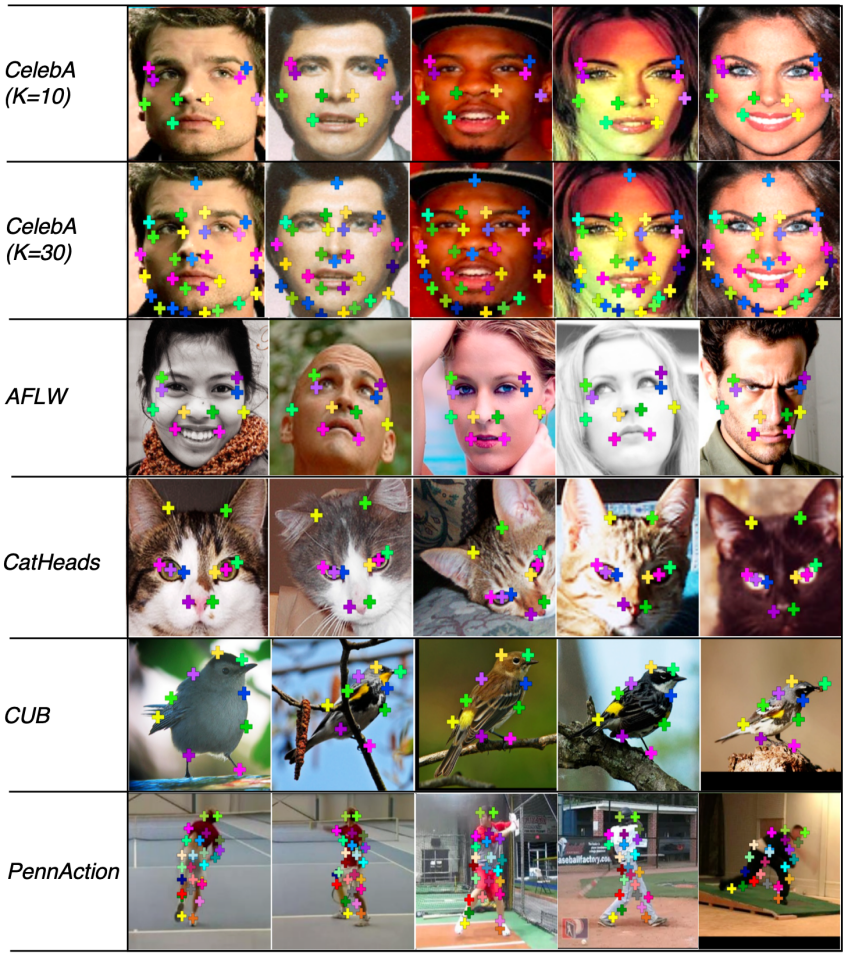


Figure 6.6: Qualitative results of our proposed approach on various object categories

Method	MAFL	AFLW
Lorenz [110] *(K=10)*	3.24	-
Shu [149]	5.45	-
Jakab et al.[72] *(K=10)*	3.19	6.86
Zhang et al. [200] *(K=10)*	3.46	7.01
Sanchez [144] *(K=10)*	3.99	6.69
Sahasrabudhe [142]	6.01	-
Ours† ($K = 13$)	4.18	7.42
Ours‡ ($K = 49$) [114]	4.12	7.37
Ours *(K=10)*	3.83	7.18

Table 6.7: **Forward-NME** on MAFL and AFLW (normalized by inter-ocular distance) evaluated over 5 groundtruth landmarks. The regression is trained with max N to be consistent with previous work. † Our approach as introduced in Chapter 4. ‡ Our approach as discussed in Chapter 5.

work (our approach detected 13 unsupervised landmarks as described in Chapter 4 and 49 unsupervised landmarks as described in Chapter 5). For cumulative curves, the error is calculated w.r.t 68-standard facial landmarks. From our results on all facial datasets (Fig. 6.7), we can see that our method provides the best overall performance (considering both forward and backward errors). Moreover, we outperform our previous method (as described in Chapter 5.2.2) in all comparison (see Fig. 5.9 of the previous Chapter, curves were not included here for clarity of visualisation).

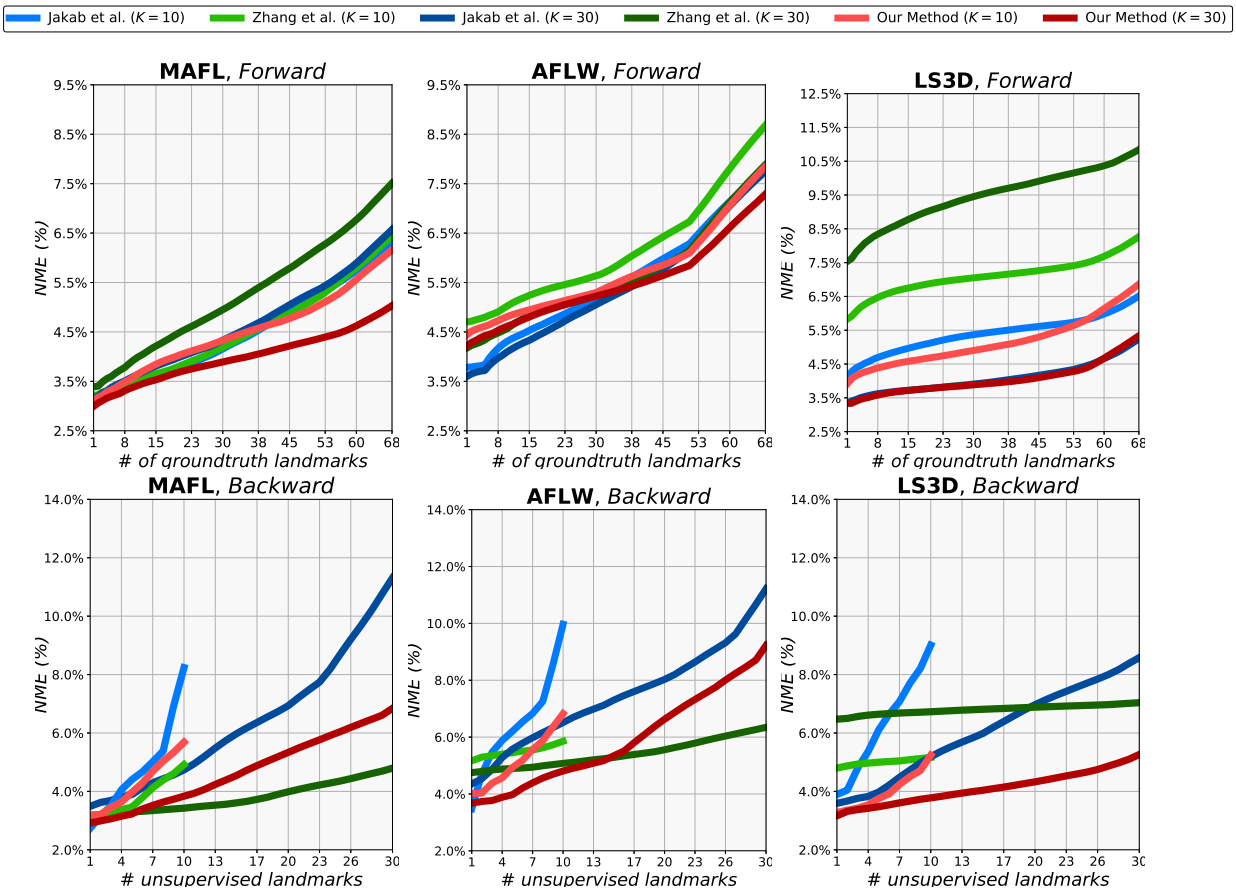


Figure 6.7: CED curves for forward and backward errors. We compare our method with [72, 200] (for $K = 10, 30$). Where possible, we used pre-trained models. Otherwise, we re-trained these methods using the publicly available code. A set of 300 training images is used to train the regressors. Error is measured w.r.t. the 68-landmark configuration typically used in face alignment.

As also observed in the previous Chapter, our approach surpasses other methods when evaluation is performed w.r.t to all 68-facial landmarks (compared to standard 5 landmark evaluation on MAFL and AFLW presented on Table 6.7 where we maintain competitive performance). The 5 facial landmark configuration includes points in uniform areas and not

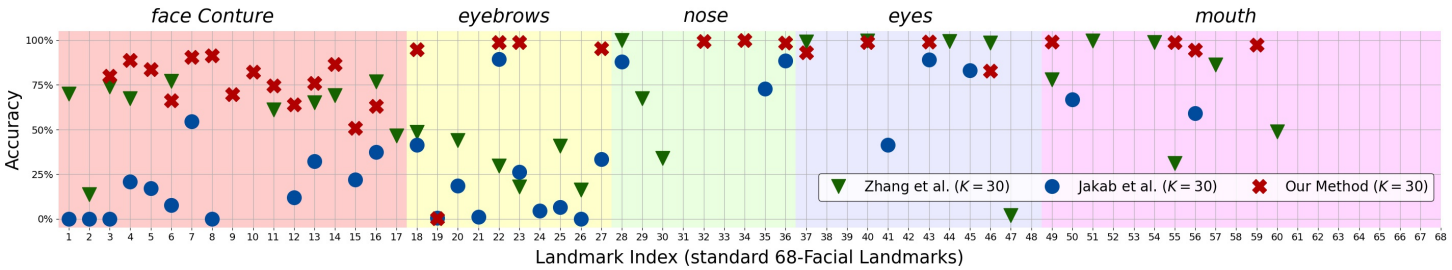


Figure 6.8: Evaluation of the ability of raw unsupervised landmarks to capture supervised landmark locations on CelebA. Each unsupervised landmark are mapped to the best corresponding supervised landmark using the Hungarian Algorithm. Then accuracy is calculated for a distance threshold of $d = 10px$. 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, e.t.c.)

CatHeads Forward-NME (%)			
Thewlis[164]	Zhang[200]	Lorenz[110]	Ours
26.94	14.84	9.30	9.31

Table 6.8: Performance on the CatHeads dataset [197]. 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).

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 this, we evaluate how accurately raw unsupervised landmarks track supervised landmark locations in Fig. 6.8. Each one 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 [72, 200] are mostly uniformly speared 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. 6.8. Our method reaches a similar error value as the best performing method of [110].

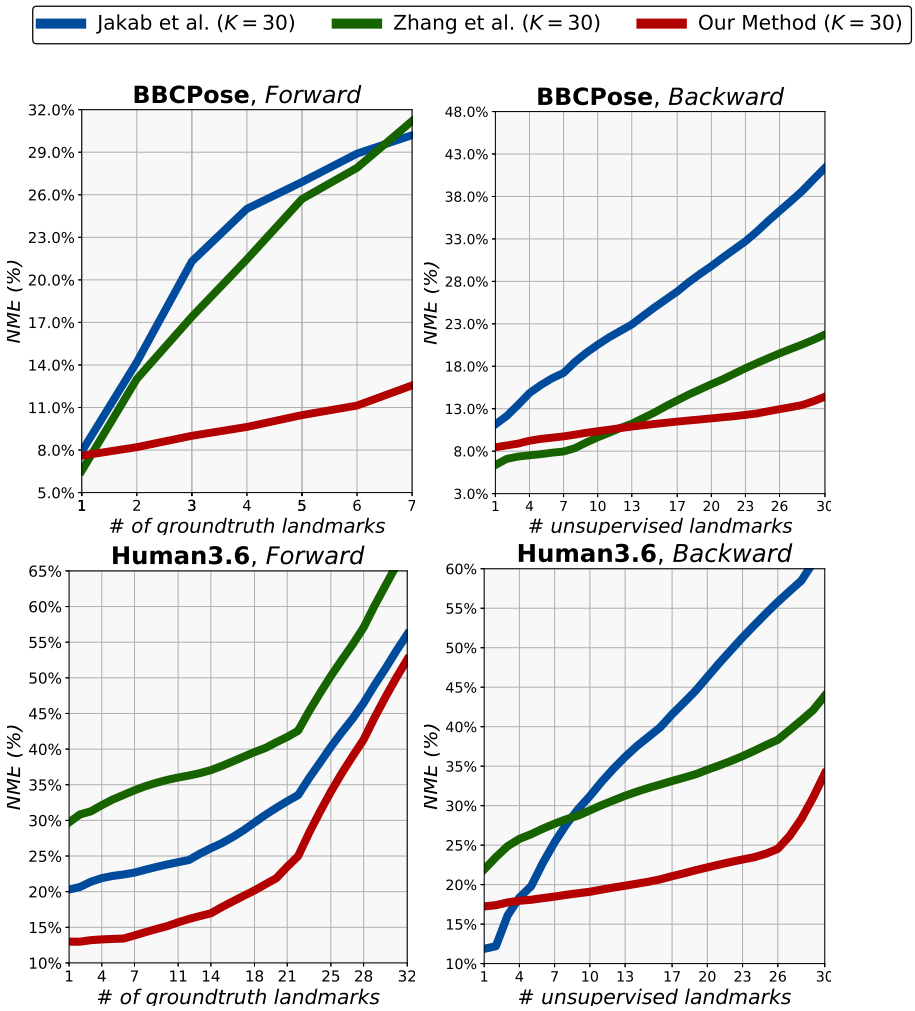


Figure 6.9: Evaluation on BBCPose and Human3.6 datasets. CED curves for the forward and backward errors, computed for a regressor trained with 800 samples. We compare our method with [72, 200] (re-trained using the publicly available code). All methods are trained to discover 30 landmarks.

6.4.2 Evaluation on human pose datasets

Performance of our method on the BBCPose and Human3.6M datasets is shown in Fig. 6.9. Note that in this experiment, all methods are trained without temporal supervision. For both datasets, our approach demonstrates significantly better performance. As also observed in the previous Chapter, the forward error in Human3.6M sharply increases for all methods when more than 22 landmarks are considered because the hands of the subject are not captured by any method.

In Table 6.9 we measure the accuracy of regressed landmarks on the BBCPose

BBCPose Regressed Landmark Accuracy (%)					
Method	Head	Shldrs	Elbws	Hands	Avg
		Supervised			
Yang [187]	63.40	53.70	49.20	46.10	51.63
Pfister [130]	74.90	53.05	46.00	71.40	59.40
Chen [28]	65.90	47.90	66.50	76.80	64.10
Charles [24]	95.40	72.95	68.70	90.30	79.90
Pfister [129]	98.00	88.45	77.10	93.50	88.01
		Unsupervised			
Jakab [72](selfsup)	81.01	49.05	53.05	70.10	60.79
Jakab [72]	76.10	56.50	70.70	74.30	68.44
Lorenz [110]	-	-	-	-	74.50
Ours	97.89	49.65	71.26	84.90	75.93

Table 6.9: Accuracy of regressed landmarks on BBCPose measured as %-age of points within $d = 6px$ from the ground-truth for an image resolution of $128px$. Results for other methods taken directly from the papers. All unsupervised methods in this experiment utilise temporal information.

PennAction Raw Landmark Accuracy (%)							
Method	Head	Shldrs	Elbws	Hands	Waist	Knees	Legs
Jakab [72]	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 6.10: 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 [72] (image resolution of $128px$). For this experiment, examined methods do not utilise temporal information. † Our method as described in Chapter 5.

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. Some visual examples of unsupervised landmarks that maximally correspond to ground-truth points are given in Fig. 6.10.

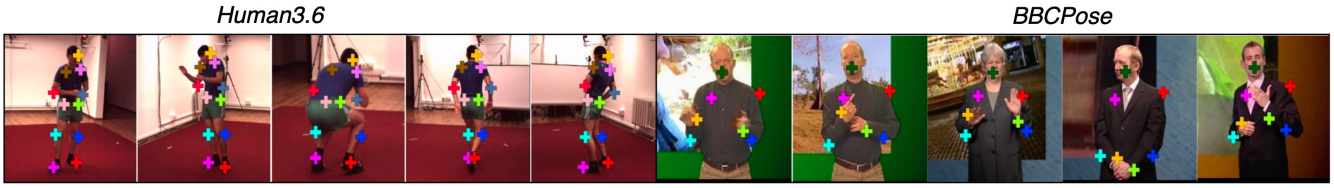


Figure 6.10: Examples on Human3.6 and BBCPose databases. We show the unsupervised landmarks that maximally corresponding to the provided groundtruth (selected through the Hungarian Algorithm).

Human3.6 Raw Landmark Evaluation

Accuracy(%)							
Method	Head	Shldrs	Elbws	Waist	Knees	Legs	Avg
Zhang [200]	20.9	53.1	51.0	43.7	85.6	2.0	42.7
Jakab [72]	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
PCK(%)							
Method	Head	Shldrs	Elbws	Waist	Knees	Legs	Avg
Zhang [200]	11.1	34.8	44.6	20.9	69.3	0.50	30.2
Jakab [72]	0.20	39.8	19.3	15.2	2.15	14.1	15.1
Ours	51.7	86.3	43.0	92.2	83.9	63.5	70.1
Average Precision and Recall (over OKS)							
Method	AP	$AP_{0.5}$	$AP_{0.4}$	AR	$AR_{0.5}$	$AR_{0.4}$	
Zhang [200]	0.02	0.17	0.67	0.06	0.41	0.82	
Jakab [72]	0.0	0.0	0.07	0.0	0.01	0.25	
Ours	0.22	0.84	0.95	0.30	0.91	0.97	

Table 6.11: Evaluation of raw discovered landmarks that correspond maximally to each ground-truth point. We report accuracy as %-age of points within $d = 6px$ from the ground-truth (image resolution of $128px$), the Percentage of Correct Keypoints (PCK) calculated over a threshold of 0.3 of torso length, as well as Average Precision (AP) and Recall (AR) (commonly used to evaluate *supervised* human pose estimation methods, for example [32]). We also calculate AP and AR with a relaxed OKS threshold of 0.4.

We also 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 (Tables 6.11, 6.10). We observe that our approach is able to 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 (including our previous work as described in Chapter 5). For Human3.6 (Table 6.11) we can see that our method surpasses state-of-the-art methods in all reported metrics. For the challenging PennAction database that includes large pose variation and complicated backgrounds, we demonstrate higher accuracy (Table 6.10), whereas [72] completely underperforms in this setting. Note that we do not use temporal supervision to train examined methods. Some qualitative results of our approach for various object categories are shown in the following figures.

6.5 Chapter Conclusion

This Chapter extends our method through several technical innovations. Most notably, we remove the progressive clustering stage of our previous work, introduce a novel negative pair selection strategy and enabling flipping augmentation for unsupervised landmark detection. The proposed effective landmark detection framework achieves improved performance compared to our previous work through a simpler training pipeline. We extend our experimental analysis to challenging datasets like PennAction where other recent methods underperform when trained without additional temporal supervision.

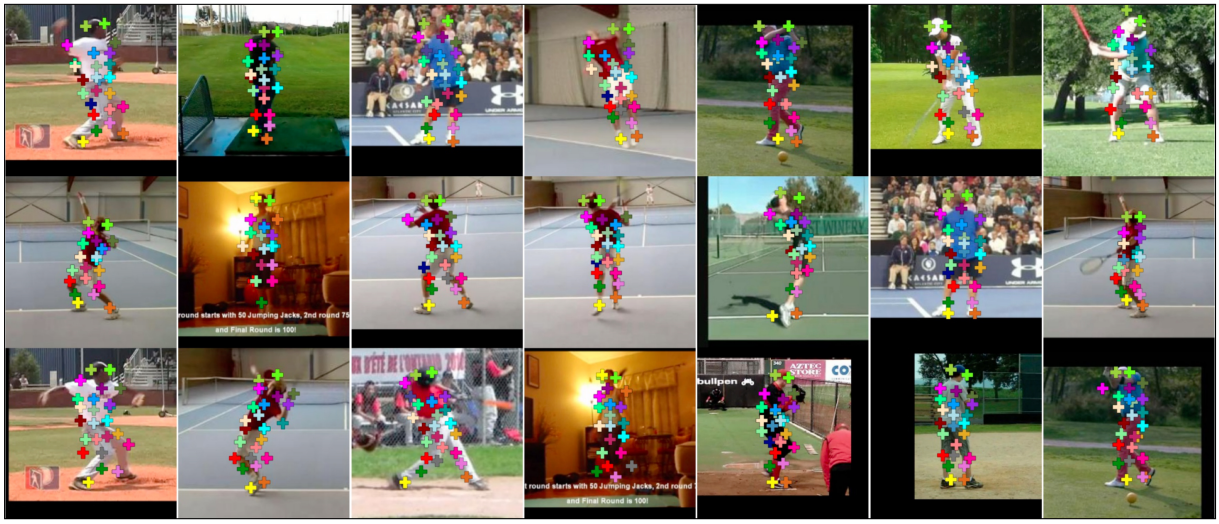


Figure 6.11: Qualitative results for our approach on PennAction.

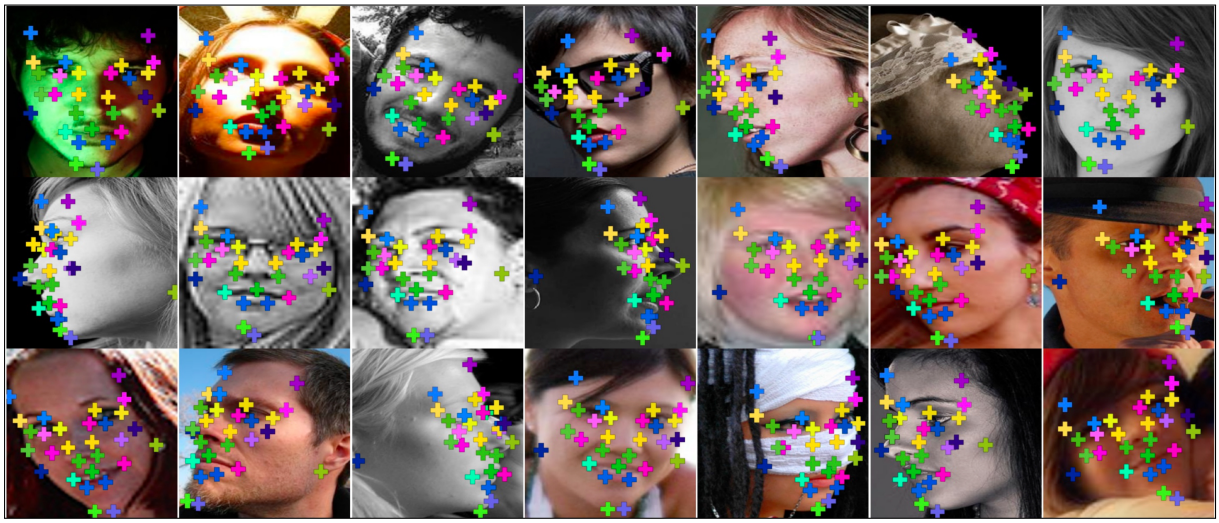


Figure 6.12: Qualitative results for our approach on LS3D.

Figure 6.13: Qualitative results for our approach on BBCPose.

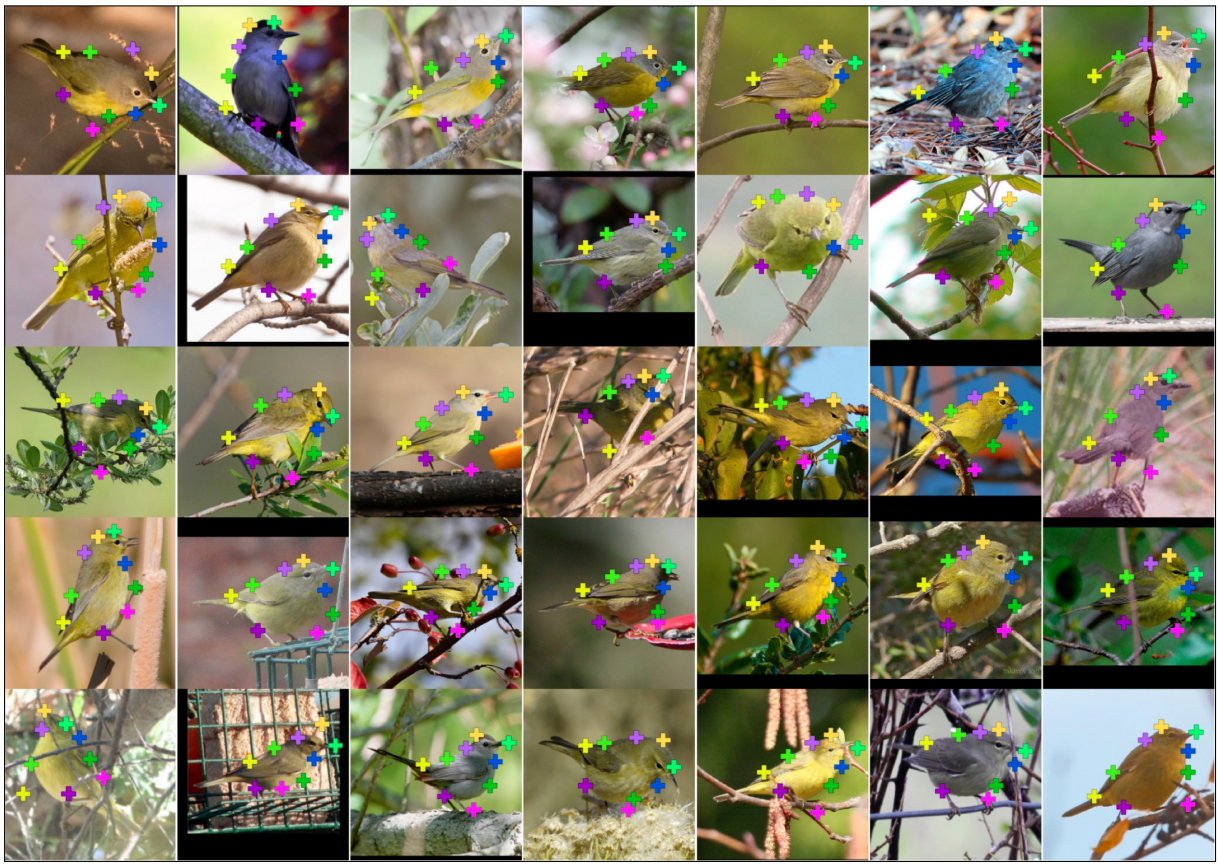


Figure 6.14: Qualitative results for our approach on CUB-200-2011.

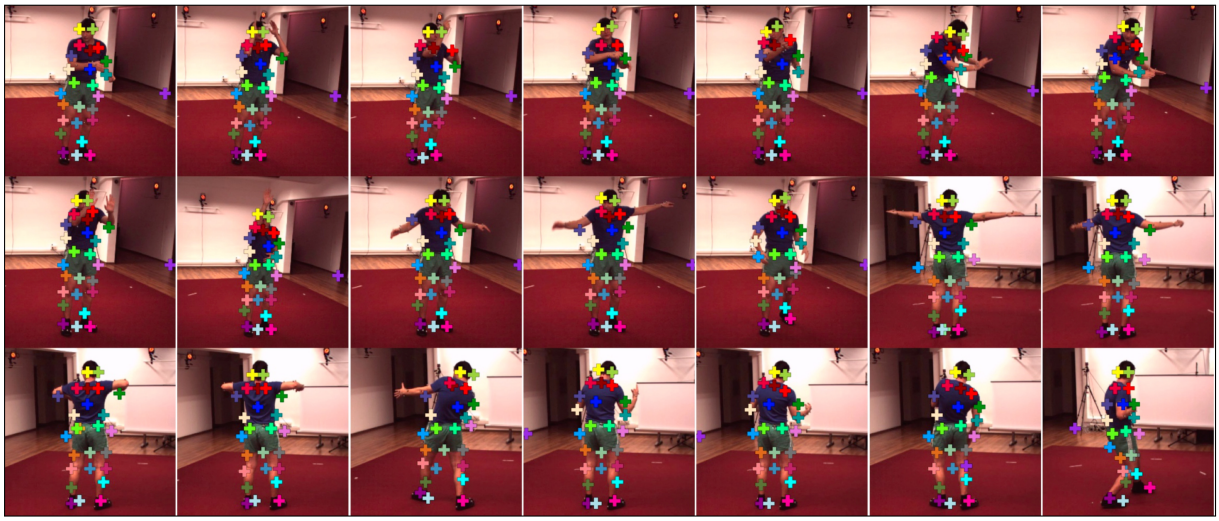


Figure 6.15: Qualitative results for our approach on Human3.6.

Figure 6.16: Qualitative results for our approach on CatHeads.

Chapter 7

Conclusions

The aim of this thesis was the development of landmark detectors for arbitrary object categories in an unsupervised manner, without requirement for expensive manual annotations. Mainly, we address two important challenges of *unsupervised landmark detection*. *(a)* The discovery of highly *semantic* landmarks. This work identifies the various analogies between keypoints and object landmarks. Through *self-training* on generic keypoints, our proposed landmark detection framework detects landmarks similar to the ones assigned from human annotators for multiple object categories (human and cat faces, human body, e.t.c). *(b)* Developing detectors that are robust to large changes in viewpoint, out-of-plane rotations and object deformations. Through *clustering correspondence*, landmark representations are learned directly from unpaired images leading to strong invariance to appearance or shape variations. Our models can fit faces even in very large poses (in the -90 to 90 range for the LS3D database) and account for object deformation (like those demonstrated by the human body in Human3.6 and PennAction databases).

Strong performance is demonstrated on highly deformable object categories, even without temporal supervision (that is necessary for related methods). We also contribute a study of training neural nets under extreme label noise for fine-grained localisation tasks and enable flipping augmentation for unsupervised landmark discovery through deep clustering. Throughout this thesis, we demonstrate state-of-the-art land-

mark detection performance on multiple challenging datasets capturing diverse object categories (LS3D, Human3.6, PennAction, BBCPose). The contributions of this thesis have also led into two a prestigious publications [114, 115] at the *NeurIPS* conference in 2020 and *IEEE's Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* in 2023. Code and pretrained models have been made publicly available on `https://github.com/malldimi1/UnsupervisedLandmarks` and `https://github.com/malldimi1/KeypointsToLandmarks` under an MIT licence.

7.1 Discussion

So far we demonstrated that the proposed approach is able to consistently discover highly semantic landmarks for various object types. However, we also note that it might not be appropriate for every possible object category. We identify and discuss various properties of object datasets that might affect the performance of our unsupervised landmark detector.

(1) Datasets with extreme interclass variation: Clustering correspondence requires some consistency in visual appearance, to produce meaningful clusters. We found that our method underperforms for object categories demonstrating extreme interclass appearance variation. An interesting example is the Shoes dataset (from UT Zappos50k [189]) commonly referenced in the unsupervised landmark discovery literature. Related methods based on image generation/reconstruction tend to perform well, since objects are depicted over a white background under uniform orientation (minimum 3D rotations), leading to strong reconstruction performance. Our approach in contrast, struggles to discover consistent clusters due to the extreme appearance variation between keypoints on different shoe types. Consider that a keypoint on the heel of a high-heel shoe has a completely different visual appearance compared to a keypoint on the back of a flip-flop, even though both points refer to the same underlying landmark location. Note that even in this case, our method would be able to discover shoe-type specific clusters but would struggle to group clusters corresponding to the same semantic landmark. A similar effect was observed

on animal datasets capturing multiple animal species with a uniform keypoint configuration like [191, 16]. Note that extreme appearance variation in animal pose estimation constitutes a challenge even for supervised methods.

(2) Object categories with landmarks on low texture areas. A basic assumption of the proposed approach is that generic keypoints would tend to overlap with object landmarks across multiple instances of the same object. Thus, the proposed approach might underperform for object categories where semantic landmarks are mostly located on flat surfaces. The SuperPoint [40] keypoint detector would not capture meaningful landmarks located on flat surfaces. It is discussed in Chapter 4, that our method underperforms when evaluated w.r.t the 5-point configuration on CelebA, comprising mostly of facial landmarks located on uniform areas of the face. In contrast, methods based on an image generation/reconstruction objective, tend to discover a large number of unsupervised landmarks in low-texture areas.

(3) Objects under full 3D rotation. We have demonstrated that our approach is able to provide invariance to large out-of-plane rotations much better compared to other recent methods. In many cases though, discovered landmarks can either be unreliable or unable to provide full invariance to large 3D rotations. In Human3.6M, many images display a full 3D rotation for humans, including poses that leave the camera behind. The visual appearance of the frontal and back view of the human body is similar and discovered landmarks cannot distinguish the left and right sides, thus full invariance to 3D rotation is not provided. To our knowledge, unsupervised landmark discovery methods are not yet able to provide complete 3D rotation invariance for human pose estimation datasets.

(4) Datasets with artefacts or repeatable background points. Since our method is bootstrapped by generic keypoints, artefacts on training images (that might fire as keypoints) or recurring patterns in the background might cause performance degradation. An example is the SimplifiedHumans, that is, the Human3.6M [69] database with the background removed through the off-the-shelf unsupervised background subtraction method (provided in the dataset). Even though related methods perform much better given a simplified background

(the reconstruction task is simplified), artefacts created by background subtraction result in a large number of noisy keypoints that might deteriorate the performance of our method. Our method, would also tend to capture repeatable keypoints as object landmarks, even when they are located in the background. This was manifested in Human3.6M database, an indoor human pose dataset captured in a single room. Multiple elements of the room (edges on doors, floor, etc) were captured as keypoints and were repeated across multiple images, leading to the discovery of unreliable background object landmarks.

Finally, even though the framework proposed in this thesis was applied to various challenging object types (faces, bodies, birds, animal faces, etc), multiple object domains and potential applications remain unexplored. We leave the adaptation of our proposed approach for unsupervised landmark discovery on novel object categories as interesting future work. Some exciting applications can be keypoint detection for medical imaging, unsupervised pose estimation for 3D objects, keypoint detection for object animation (as in [150]) or even keypoint detection for plant or root health classification (for example [7]).

7.2 Future work

The main focus of this work is the discovery of unsupervised landmarks on various object categories without manual supervision. Even though significant progress is demonstrated through this thesis, the task of *unsupervised landmark discovery* remains an open problem with significant room for improvement. We identify multiple interesting directions for future work.

Dependency to SuperPoint

Our proposed framework is bootstrapped by generic keypoints. We find that SuperPoint [40] provides the strongest initialisation for our method. Currently, SuperPoint is used as an off-the-shelf external component. However, a more elegant solution could be to perform

keypoint detection inside our proposed framework directly. One possible approach could be to infer possible keypoint locations by exploiting the information within our feature extractor as in [166]. This could simplify our proposed framework and further increase the semantic value of detected landmarks.

Domain Adaptation

The main focus of this thesis is the development of robust landmark detectors for arbitrary object categories. An interesting future direction could be following a domain adaptation paradigm. We assume that a pre-trained landmark detector, trained in a fully supervised manner for a source object category, is available. As in [144], we can adapt the pre-trained detector on the target object category simply by learning two projection matrices, from the core pretrained network (that remains frozen during training) to our detector and feature extractor heads (see subsection 5.2). This architecture could significantly reduce the number of learned parameters and potentially enhance performance since strong invariance to viewpoint variations and natural object deformations are already encoded to the pre-trained network.

Single Stage Pipeline

Our self-training framework for unsupervised landmark discovery comprises of two stages (Chapters 5,6). The first stage aims to establish landmark correspondence and recover object landmark locations. Then, pseudo labels formed from the first stage are used to train a landmark detector. An interesting future direction could be to modify our proposed framework into a single-stage pipeline. This could be achieved by including an third output head $\mathbf{\Phi}_{h,3}$ (see subsection 5.2) with $\mathcal{X} \rightarrow \mathcal{E}$, where $\mathcal{E} \in \mathbb{R}^{H_o \times W_o \times K}$ that is trained to with standard heatmap regression to detect K object landmarks. Pseudolabels can be formed during iterative self-training of stage 1 by applying K-means directly with K clusters (as in subsection 5.2.2). Special care needs to be taken to account for varying cluster order

for subsequent K-mean rounds. One solution could be to train the last Conv layer of $\mathbf{\Phi}_{h,3}$ from scratch on each self-training round.

Advancements in SSL

This thesis proposes an SSL-based approach to discover correspondences through deep clustering of landmark representations. Given the rapid progress of *self-supervised learning*, several new techniques are recently proposed that claim enhanced representation learning performance. Possible directions to capitalise on recent progress could be to increase the number of negative pairs for contrastive learning (through momentum contrast [62], larger batch sizes[25] or memory banks [178]), form informative, positive pairs by applying strong augmentation strategies (for example, RandAugment [37]), or discarding negative pairs as in [58].

Bibliography

[1] Achille, A., and Soatto, S. Emergence of invariance and disentanglement in deep representations. *ITA Workshop* (2018).

[2] Agarwal, S., Snavely, N., Seitz, S. M., and Szeliski, R. Bundle adjustment in the large. In *ECCV* (2010).

[3] Agrawal, M., Konolige, K., and Blas, M. R. Censure: Center surround extremas for realtime feature detection and matching. In *ECCV* (2008).

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

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

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

[7] Azimi, S., Lall, B., and Gandhi, T. K. Performance evalution of 3d keypoint detectors and descriptors for plants health classification. *MVA* (2019).

[8] Bailo, O., Rameau, F., Joo, K., Park, J., Bogdan, O., and Kweon, I. S. Efficient adaptive non-maximal suppression algorithms for homogeneous spatial keypoint distribution. *Pattern Recognition Letters 106* (2018).

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

[10] Bookstein, F. L. Principal warps: Thin-plate splines and the decomposition of deformations. *TPAMI* (1989).

[11] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T. J., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. *NeurIPS* (2020).

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

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

[14] Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I. D., and Leonard, J. J. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics 32* (2016), 1309–1332.

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

[16] Cao, J., Tang, H., Fang, H.-S., Shen, X., Lu, C., and Tai, Y.-W. Cross-domain adaptation for animal pose estimation. In *ICCV* (2019).

[17] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. *CVPR* (2016), 1302–1310.

[18] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. *CVPR* (2017).

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

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

[21] Carreira, J., Agrawal, P., Fragkiadaki, K., and Malik, J. Human pose estimation with iterative error feedback. *CVPR* (2016), 4733–4742.

[22] Cascante-Bonilla, P., Tan, F., Qi, Y., and Ordonez, V. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *AAAI* (2021).

[23] Chang, A. X., Funkhouser, T. A., Guibas, L. J., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., and Yu, F. Shapenet: An information-rich 3d model repository. *ArXiv* (2015).

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

[25] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. *PMLR* (2020).

[26] Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. Big self-supervised models are strong semi-supervised learners. *NeurIPS* (2020).

[27] Chen, X., and He, K. Exploring simple siamese representation learning. *CVPR* (2021), 15745–15753.

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

[29] Chen, Y., Shen, C., Wei, X.-S., Liu, L., and Yang, J. Adversarial posenet: A structure-aware convolutional network for human pose estimation. *ICCV* (2017).

[30] Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., and Sun, J. Cascaded pyramid network for multi-person pose estimation. *CVPR* (2018).

[31] Chen, Y., Wei, C., Kumar, A., and Ma, T. Self-training avoids using spurious features under domain shift. *ArXiv* (2020).

[32] Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T. S., and Zhang, L. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. *CVPR* (2020).

[33] Cheng, Z., Su, J.-C., and Maji, S. Unsupervised discovery of object landmarks via contrastive learning. *ArXiv* (2020).

[34] Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A. L., and Wang, X. Multi-context attention for human pose estimation. *CVPR* (2017).

[35] Clark, C., Yatskar, M., and Zettlemoyer, L. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *ArXiv* (2019).

[36] Cootes, T., Edwards, G. J., and Taylor, C. J. Active appearance models. *TPAMI* (2001).

[37] Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical data augmentation with no separate search. *CVPR* (2019).

[38] Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. *CVPRW* (2020).

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

[40] DeTone, D., Malisiewicz, T., and Rabinovich, A. Superpoint: Self-supervised interest point detection and description. *CVPRW* (2018).

[41] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL* (2019).

[42] Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *ICLR* (2017).

[43] Dollár, P., Welinder, P., and Perona, P. Cascaded pose regression. *CVPR* (2010).

[44] Donahue, J., Krähenbühl, P., and Darrell, T. Adversarial feature learning. *ICLR* (2017).

[45] Donahue, J., and Simonyan, K. Large scale adversarial representation learning. In *NeurIPS* (2019).

[46] Eickhoff, C. Cognitive biases in crowdsourcing. *WSDM* (2018).

[47] Eickhoff, C., Harris, C. G., de Vries, A. P., and Srinivasan, P. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *SIGIR* (2012).

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

[49] Felzenszwalb, P. F., Girshick, R. B., McAllester, D. A., and Ramanan, D. Object detection with discriminatively trained part based models. *TPAMI* (2009).

[50] Feng, Z., Kittler, J., Awais, M., Huber, P., and Wu, X. Face detection, bounding box aggregation and pose estimation for robust facial landmark localisation in the wild. *CVPRW* (2017).

[51] Fischler, M. A., and Elschlager, R. A. The representation and matching of pictorial structures. *IEEE Transactions on Computers C-22* (1973), 67–92.

[52] Ganin, Y., and Lempitsky, V. S. Unsupervised domain adaptation by backpropagation. *ICML* (2014).

[53] Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. *ICLR* (2018).

[54] Goyal, P., Caron, M., Lefaudeux, B., Xu, M., Wang, P., Pai, V., Singh, M., Liptchinsky, V., Misra, I., Joulin, A., and Bojanowski, P. Self-supervised pretraining of visual features in the wild. *ArXiv* (2021).

[55] Goyal, P., Dollár, P., Girshick, R. B., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *ArXiv* (2017).

[56] Goyal, P., Mahajan, D. K., Gupta, A., and Misra, I. Scaling and benchmarking self-supervised visual representation learning. *ICCV* (2019).

[57] Graves, A. Generating sequences with recurrent neural networks. *ArXiv* (2013).

[58] Grill, J.-B., Strub, F., Altch'e, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. Á., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent: A new approach to self-supervised learning. *NeurIPS abs/2006.07733* (2020).

[59] Grover, A., and Leskovec, J. node2vec: Scalable feature learning for networks. *SIGKDD* (2016).

[60] Güler, R. A., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S., and Kokkinos, I. Densereg: Fully convolutional dense shape regression in-the-wild. *CVPR* (2017).

[61] Harris, C. G., and Stephens, M. J. A combined corner and edge detector. In *Alvey Vision Conference* (1988).

[62] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. B. Momentum contrast for unsupervised visual representation learning. *CVPR* (2020), 9726–9735.

[63] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *CVPR* (2016).

[64] Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *ICLR* (2019).

[65] Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K. Q. Deep networks with stochastic depth. In *ECCV* (2016).

[66] Huang, J., Zhu, Z., Guo, F., and Huang, G. The devil is in the details: Delving into unbiased data processing for human pose estimation. *CVPR* (2020), 5699–5708.

[67] Hummel, H. I., Pessanha, F., Salah, A. A., van Loon, T., and Veltkamp, R. C. Automatic pain detection on horse and donkey faces. *FG* (2020).

[68] Ioffe, S., and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML* (2015).

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

[70] Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. Spatial transformer networks. In *NIPS* (2015).

[71] Jain, V., and Learned-Miller, E. G. Fddb: A benchmark for face detection in unconstrained settings.

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

[73] Jakab, T., Gupta, A., Bilen, H., and Vedaldi, A. Learning landmarks from unaligned data using image translation. *ArXiv* (2019).

[74] Jakab, T., Gupta, A., Bilen, H., and Vedaldi, A. Self-supervised learning of interpretable keypoints from unlabelled videos. *CVPR* (2020).

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

[76] Jing, J., Gao, T., Zhang, W., Gao, Y., and Sun, C. Image feature information extraction for interest point detection: A comprehensive review. *ArXiv* (2021).

[77] Jing, L., and Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. *TPAMI 43* (2021), 4037–4058.

[78] Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with gpus. *ArXiv* (2017).

[79] Khan, M. H., McDonagh, J., Khan, S. H., Shahabuddin, M., Arora, A., Khan, F. S., Shao, L., and Tzimiropoulos, G. Animalweb: A large-scale hierarchical dataset of annotated animal faces. *CVPR* (2020).

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

[81] Kim, D., Cho, D., Yoo, D., and Kweon, I.-S. Learning image representations by completing damaged jigsaw puzzles. *WACV* (2018), 793–802.

[82] Kim, Y., Nam, S., Cho, I., and Kim, S. J. Unsupervised keypoint learning for guiding class-conditional video prediction. In *NeurIPS* (2019).

[83] Kingma, D. P., and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS* (2018).

[84] Kipf, T., and Welling, M. Variational graph auto-encoders. *ArXiv* (2016).

[85] Koch, G., Zemel, R., Salakhutdinov, R., et al. Siamese neural networks for one-shot image recognition. In *ICML workshop* (2015).

[86] Kong, L., de Masson d'Autume, C., Ling, W., Yu, L., Dai, Z., and Yogatama, D. A mutual information maximization perspective of language representation learning. *ICLR* (2020).

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

[88] Krizhevsky, A. One weird trick for parallelizing convolutional neural networks. *ArXiv* (2014).

[89] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM 60* (2012).

[90] Kuhn, H. W. The hungarian method for the assignment problem.

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

[92] Lai, H., Xiao, S., Pan, Y., Cui, Z., Feng, J., Xu, C., Yin, J., and Yan, S. Deep recurrent regression for facial landmark detection. *TCSVT* (2018).

[93] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *ArXiv* (2020).

[94] Larsson, G., Maire, M., and Shakhnarovich, G. Learning representations for automatic colorization. *ECCV* (2016).

[95] Ledig, C., Theis, L., Huszár, F., Caballero, J., Aitken, A. P., Tejani, A., Totz, J., Wang, Z., and Shi, W. Photo-realistic single image super-resolution using a generative adversarial network. *CVPR* (2017).

[96] Lee, D.-H. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks.

[97] Lenc, K., and Vedaldi, A. Understanding image representations by measuring their equivariance and equivalence. *CVPR* (2015).

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

[99] Leutenegger, S., Chli, M., and Siegwart, R. Y. Brisk: Binary robust invariant scalable keypoints. *ICCV* (2011).

[100] Li, B., Xiao, R., Li, Z., Cai, R., Lu, B.-L., and Zhang, L. Rank-sift: Learning to rank repeatable local interest points. *CVPR* (2011).

[101] Li, C., and Lee, G. H. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. *CVPR* (2021).

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

[103] Li, W., Liao, H., Miao, S., Lu, L., and Luo, J. Unsupervised learning of facial landmarks based on inter-intra subject consistencies. *ICPR* (2021).

[104] Liao, R., Schwing, A. G., Zemel, R. S., and Urtasun, R. Learning deep parsimonious representations. In *NIPS* (2016).

[105] Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV* (2014).

[106] Lindeberg, T. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. *IJCV 11* (2005), 283–318.

[107] Liu, X., Zhang, F., Hou, Z., Wang, Z., Mian, L., Zhang, J., and Tang, J. Self-supervised learning: Generative or contrastive. *ArXiv* (2020).

[108] Liu, Z., Luo, P., Qiu, S., Wang, X., and Tang, X. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR* (2016).

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

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

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

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

[113] Luvizon, D. C., Tabia, H., and Picard, D. Human pose regression by combining indirect part detection and contextual information. *ArXiv* (2019).

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

[115] Mallis, D., Sanchez, E., Bell, M., and Tzimiropoulos, G. From keypoints to object landmarks via self-training correspondence: A novel approach to unsupervised landmark discovery. *ArXiv* (2022).

[116] Marimón, D., Bonnin, A., Adamek, T., and Gimeno, R. Darts: Efficient scale-space extraction of daisy keypoints. *CVPR* (2010).

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

[118] Mirza, M., and Osindero, S. Conditional generative adversarial nets. *ArXiv* (2014).

[119] Misra, I., and van der Maaten, L. Self-supervised learning of pretext-invariant representations. *CVPR* (2020), 6706–6716.

[120] Mokhtarian, F., and Suomela, R. Robust image corner detection through curvature scale space. *TPAMI* (1998).

[121] Myronenko, A., and Song, X. B. Point set registration: Coherent point drift. *TPAMI* (2010).

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

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

[124] Nibali, A., He, Z., Morgan, S., and Prendergast, L. Numerical coordinate regression with convolutional neural networks. *ArXiv* (2018).

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

[126] Oczak, M., Maschat, K., Berckmans, D., Vranken, E., and Baumgartner, J. Automatic estimation of number of piglets in a pen during farrowing, using image analysis. *Biosystems Engineering* (2016).

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

[128] Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., and Efros, A. A. Context encoders: Feature learning by inpainting. *CVPR* (2016), 2536–2544.

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

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

[131] Radford, A., and Narasimhan, K. Improving language understanding by generative pre-training.

[132] Radosavovic, I., Dollár, P., Girshick, R. B., Gkioxari, G., and He, K. Data distillation: Towards omni-supervised learning. *CVPR* (2017).

[133] Radosavovic, I., Dollár, P., Girshick, R. B., Gkioxari, G., and He, K. Data distillation: Towards omni-supervised learning. *CVPR* (2018).

[134] Ramakrishna, V., Munoz, D., Hebert, M., Bagnell, J. A., and Sheikh, Y. Pose machines: Articulated pose estimation via inference machines. In *ECCV* (2014).

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

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

[137] Rolnick, D., Veit, A., Belongie, S. J., and Shavit, N. Deep learning is robust to massive label noise. *ArXiv* (2017).

[138] Rosten, E., and Drummond, T. Machine learning for high-speed corner detection. In *ECCV* (2006).

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

[140] Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *ArXiv* (2019).

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

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

[143] Salti, S., Lanza, A., and di Stefano, L. Keypoints from symmetries by wave propagation. *CVPR* (2013), 2898–2905.

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

[145] Savinov, N., Seki, A., Ladicky, L., Sattler, T., and Pollefeys, M. Quad-networks: Unsupervised learning to rank for interest point detection. *CVPR* (2017).

[146] Schneider, S., Baevski, A., Collobert, R., and Auli, M. wav2vec: Unsupervised pre-training for speech recognition. In *INTERSPEECH* (2019).

[147] Shen, J., Zafeiriou, S., Chrysos, G. G., Kossaifi, J., Tzimiropoulos, G., and Pantic, M. The first facial landmark tracking in-the-wild challenge: Benchmark and results. *ICCVW* (2015), 1003–1011.

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

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

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

[151] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556* (2015).

[152] Smith, S. M., and Brady, M. Susan—a new approach to low level image processing. *IJCV* (2004).

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

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

[155] Sun, K., Xiao, B., Liu, D., and Wang, J. Deep high-resolution representation learning for human pose estimation. In *CVPR* (2019).

[156] Sun, X., Shang, J., Liang, S., and Wei, Y. Compositional human pose regression. *ICCV* (2018).

[157] Sun, Y., Wang, X., and Tang, X. Deep convolutional network cascade for facial point detection. *CVPR* (2013).

[158] Sutskever, I., Hinton, G. E., and Taylor, G. W. The recurrent temporal restricted boltzmann machine. In *NIPS* (2008).

[159] Suwajanakorn, S., Snavely, N., Tompson, J., and Norouzi, M. Discovery of latent 3d keypoints via end-to-end geometric reasoning. In *NeurIPS* (2018).

[160] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. *CVPR* (2015).

[161] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. *ICLR* (2014).

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

[163] Thewlis, J., Bilen, H., and Vedaldi, A. Unsupervised learning of object frames by dense equivariant image labelling.

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

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

[166] Tian, Y., Balntas, V., Ng, T., Laguna, A. B., Demiris, Y., and Mikolajczyk, K. D2d: Keypoint extraction with describe to detect approach. In *ACCV* (2020).

[167] Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning. *NeurIPS* (2020).

[168] Tompson, J., Goroshin, R., Jain, A., LeCun, Y., and Bregler, C. Efficient object localization using convolutional networks. *CVPR* (2015), 648–656.

[169] Tompson, J., Jain, A., LeCun, Y., and Bregler, C. Joint training of a convolutional network and a graphical model for human pose estimation. *NIPS* (2014).

[170] Toshev, A., and Szegedy, C. Deeppose: Human pose estimation via deep neural networks. *CVPR* (2014).

[171] Tu, L., Lalwani, G., Gella, S., and He, H. An empirical study on robustness to spurious correlations using pre-trained language models. *TACL* (2020).

[172] Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. *CVPR* (2017).

[173] van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *ArXiv* (2018).

[174] van der Maaten, L., and Hinton, G. E. Visualizing data using t-sne.

[175] Verdie, Y., Yi, K. M., Fua, P., and Lepetit, V. Tilde: A temporally invariant learned detector. *CVPR* (2015).

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

[177] Wu, Y., and He, K. Group normalization. In *ECCV* (2018).

[178] Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. *CVPR* (2018), 3733–3742.

[179] Xia, G.-S., Delon, J., and Gousseau, Y. Accurate junction detection and characterization in natural images. *IJCV* (2013).

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

[181] Xie, J., Girshick, R. B., and Farhadi, A. Unsupervised deep embedding for clustering analysis. In *ICML* (2015).

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

[183] Xie, Q., Hovy, E. H., Luong, M.-T., and Le, Q. V. Self-training with noisy student improves imagenet classification. *CVPR* (2020), 10684–10695.

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

[185] Yang, J., Parikh, D., and Batra, D. Joint unsupervised learning of deep representations and image clusters. *CVPR* (2016).

[186] Yang, W., Li, S., Ouyang, W., Li, H., and Wang, X. Learning feature pyramids for human pose estimation. *ICCV* (2017).

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

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

[189] Yu, A., and Grauman, K. Fine-grained visual comparisons with local learning. *CVPR* (2014).

[190] Yu, C., Xiao, B., Gao, C., Yuan, L., Zhang, L., Sang, N., and Wang, J. Lite-hrnet: A lightweight high-resolution network. *CVPR* (2021).

[191] Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., and Tao, D. Ap-10k: A benchmark for animal pose estimation in the wild. *ArXiv* (2021).

[192] Zeiler, M. D., and Fergus, R. Visualizing and understanding convolutional networks. In *ECCV* (2014).

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

[194] Zhang, F., Zhu, X., Dai, H., Ye, M., and Zhu, C. Distribution-aware coordinate representation for human pose estimation. *CVPR* (2020).

[195] Zhang, F., Zhu, X., and Ye, M. Fast human pose estimation. *CVPR* (2019).

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

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

[198] Zhang, W., Zhu, M., and Derpanis, K. From actemes to action: A strongly-supervised representation for detailed action understanding. *ICCV* (2013), 2248–2255.

[199] Zhang, X., Yu, F. X., Karaman, S., and Chang, S.-F. Learning discriminative and transformation covariant local feature detectors. *CVPR* (2017).

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

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

[202] Zhang, Z., Luo, P., Loy, C. C., and Tang, X. Learning deep representation for face alignment with auxiliary attributes. *TPAMI 38* (2016), 918–930.

[203] Zheng, C., Wu, W., Yang, T., Zhu, S., Chen, C., Liu, R., Shen, J., Kehtarnavaz, N., and Shah, M. Deep learning-based human pose estimation: A survey. *ArXiv* (2020).

[204] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *ICCV* (2017).

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

[206] Zhu, X., and Wu, X. Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review 22* (2003), 177–210.

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

[208] Zoph, B., Ghiasi, G., Lin, T.-Y., Cui, Y., Liu, H., Cubuk, E. D., and Le, Q. V. Rethinking pre-training and self-training. *NeurIPS* (2020).