Unsupervised Learning of Landmarks by Descriptor Vector Exchange

James Thewlis*  
Unitary  
james@unitary.ai

Samuel Albanie*  
VGG, University of Oxford  
albanie@robots.ox.ac.uk

Hakan Bilen  
University of Edinburgh  
hbilen@ed.ac.uk

Andrea Vedaldi  
VGG, University of Oxford  
vedaldi@robots.ox.ac.uk

Figure 1: We propose Descriptor Vector Exchange (DVE), a mechanism that enables unsupervised learning of robust high-dimensional dense embeddings with equivariance losses. The embeddings learned for the category of faces are visualised in the figure above with the help of a query image [8], shown in the centre of the figure. (Left): We colour the locations of pixel embeddings that form the nearest neighbours of the query reference points. (Right): The same reference points are used to retrieve patches amongst a collection of face images. The result is an approximate face mosaic, matching parts across different identities despite the fact that no landmark annotations of any kind were used during learning.

Abstract

Equivariance to random image transformations is an effective method to learn landmarks of object categories, such as the eyes and the nose in faces, without manual supervision. However, this method does not explicitly guarantee that the learned landmarks are consistent with changes between different instances of the same object, such as different facial identities. In this paper, we develop a new perspective on the equivariance approach by noting that dense landmark detectors can be interpreted as local image descriptors equipped with invariance to intra-category variations. We then propose a direct method to enforce such an invariance in the standard equivariant loss. We do so by exchanging descriptor vectors between images of different object instances prior to matching them geometrically. In this manner, the same vectors must work regardless of the specific object identity considered. We use this approach to learn vectors that can simultaneously be interpreted as local descriptors and dense landmarks, combining the advantages of both. Experiments on standard benchmarks show that this approach can match, and in some cases surpass state-of-the-art performance amongst existing methods that learn landmarks without supervision. Code is available at www.robots.ox.ac.uk/~vgg/research/DVE/.

1. Introduction

Learning without manual supervision remains an open problem in machine learning and computer vision. Even recent advances in self-supervision [15, 17] are often limited to learning generic feature extractors and still require some manually annotated data to solve a concrete task such as landmark detection. In this paper, we thus consider the problem of learning the landmarks of an object category, such as the eyes and nose in faces, without any manual annotation. Namely, given as input a collection of images of a certain object, such as images of faces, the goal is to learn what landmarks exist and how to detect them.

In the absence of manual annotations, an alternative supervisory signal is required. Recently, [46] proposed to
build on the fact that landmark detectors are equivariant to image transformations. For example, if one translates or rotates a face, then the locations of the eyes and nose follow suit. Equivariance can be used as a learning signal by applying random synthetic warps to images of an object and then requiring the landmark detector to be consistent with these transformations.

The main weakness of this approach is that equivariance can only be imposed for transformations of specific images. This means that a landmark detector can be perfectly consistent with transformation applied to a specific face and still match an eye in a person and the nose in another. In this approach, achieving consistency across object instances is left to the generalisation capabilities of the underlying learning algorithm.

In this paper, we offer a new perspective on the problem of learning landmarks, generalising prior work and addressing its shortcomings. We start by establishing a link between two apparently distinct concepts: landmarks and local image descriptors (fig. 2). Recall that a descriptor, such as SIFT, is a vector describing the appearance of the image around a given point. Descriptors can establish correspondences between images because they are invariant to viewing effects such as viewpoint changes. However, similar to descriptors, landmarks can also establish image correspondences by matching concepts such as eyes or noses detected in different images.

Thus invariant descriptors and landmark detectors are similar, but landmarks are invariant to intra-class variations in addition to viewing effects. We can make this analogy precise if we consider dense descriptors and landmarks [45, 9, 42]. A dense descriptor associates to each image pixel a C-dimensional vector, whereas a dense landmark detector associates to each pixel a 2D vector, which is the index of the landmark in a (u, v) parameterisation of the object surface. Thus we can interpret a landmark as a tiny 2D descriptor. Due to its small dimensionality, a landmark loses the ability to encode instance-specific details of the appearance, but gains robustness to intra-class variations.

Generalising this idea, we note that any invariant descriptor can be turned into a landmark detector by equipping it with robustness to intra-class variations. Here we propose a new method that can do so without reducing the dimensionality of the descriptor vectors. The formulation still considers pairs of synthetically-transformed images as [45] do, but this time landmarks are represented by arbitrary C-dimensional vectors. Then, before geometric consistency (equivariance) is enforced, the landmark vectors extracted from one image are exchanged with similar vectors extracted from other random images of the object. This way geometric consistency between an image and its transformations can only be achieved if vectors have an intra-class validity, and thus effectively characterise landmarks.

![Figure 2: Descriptor-landmark hierarchy. A local invariant descriptor maps image pixels to distinctive vectors that are invariant to viewing conditions such as a viewpoint. A dense landmark detector maps pixels to unique points of the object’s surface, such as eyes and nose in faces, to points on the surface of a sphere. Both produce invariant and distinctive vectors, but landmarks are also invariant to intra-class variations. Taken together, they represent a hierarchy of distinctive pixel embeddings of increasing invariance.](image)

Empirically (section 4), we show that the key advantage of this formulation, which we term Descriptor Vector Exchange (DVE), is that it produces embedding vectors that simultaneously work well as instance-specific image descriptors and landmarks, capturing in a single representation the advantages of both, and validating our intuition.

2. Related work

**General image matching.** Image matching based on local features has been an extensively studied problem in the literature with applications to wide-baseline stereo matching [38] and image retrieval [48]. The generic pipeline contains the following steps: i) detecting a sparse set of interest points [28] that are covariant with a class of transformations, ii) extracting local descriptors (e.g. [27, 47]) at these points that are invariant to viewpoint and illumination changes, and iii) matching the nearest neighbour descriptors across images with an optional geometric verification. While the majority of the image matching methods rely on hand-crafted detectors and descriptors, recent work show that CNN-based models can successfully be trained to detect covariant detectors [23] and invariant descriptors [52, 36]. We build our method on similar principles, covariance and invariance, but with an important difference that it can learn intrinsic features for object categories in contrast to generic ones.

**Cross-instance object matching.** The SIFT Flow
method [24] extends the problem of finding dense correspondences between same object instances to different instances by matching their SIFT features [27] in a variational framework. This work is further improved by using multi-scale patches [11], establishing region correspondences [10] and replacing SIFT features with CNN ones [26]. In addition, Learned-Miller [21] generalises the dense correspondences between image pairs to an arbitrary number of images by continuously warping each image via a parametric transformation. RSA [37], Collection Flow [18] and Mobahi et al. [29] project a collection of images into a lower dimensional subspace and perform a joint alignment among the projected images. AnchorNet [34] learns semantically meaningful parts across categories, although is trained with image labels.

**Transitivity.** The use of transitivity to regularise structured data has been proposed by several authors [44, 51, 57, 58] in the literature. Earlier examples [44, 51] employ this principle to achieve forward-backward consistency in object tracking and to identify inconsistent geometric relations in structure from motion respectively. Zhou et al. [57, 58] enforce a geometric consistency to jointly align image sets or supervise deep neural networks in dense semantic alignment by establishing a cycle between each image pair and a 3D CAD model. DVE also builds on the same general principle of transitivity, however, it operates in the space of appearance embeddings in contrast to verification of subsequent image warps to a composition.

**Unsupervised learning of object structure.** Visual object characterisation (e.g. [3, 7, 22, 4, 5]) has a long history in computer vision with extensive work in facial landmark detection and human body pose estimation. A recent unsupervised method that can learn geometric transformations to optimise classification accuracy is the spatial transformer network [12]. However, this method does not learn any explicit object geometry. Similarly, WarpNet [16] and geometric matching networks [39] train neural networks to predict relative transformations between image pairs. These methods are limited to perform only on image pairs and do not learn an invariant geometric embedding for the object. Most related to our work, [46] characterises objects by learning landmarks that are consistent with geometric transformations without any manual supervision, while [33] similarly use such transformations for semantic matching. The authors of [46] extended their approach to extract a dense set of landmarks by projecting the raw pixels on a surface in [45]. Similar work [41] leverages frame-to-frame correspondence using Dynamic Fusion [31] as supervision to learn a dense labelling for human images. We build our method, DVE, on these approaches and further extend them in significant ways. First, we learn more versatile descriptors that can encode both generic and object-specific landmarks and show that we can gradually learn to move from generic to specific ones. Second, we improve the cross-instance generalisation ability by better regularising the embedding space with the use of transitivity. Finally, we show that DVE both qualitatively and quantitatively outperforms [46, 45] in facial landmark detection (section 4). Recent work [54, 13, 49, 42] proposes to disentangle appearance from pose by estimating dense deformation field [49, 42] and by learning landmark positions to reconstruct one sample from another. We compare DVE to these approaches in section 4.

**3. Method**

We first summarise the method of [45] and then introduce DVE, our extension to their approach.

**3.1. Learning dense landmarks using equivariance**

Denote by \( x \in \mathbb{R}^{3 \times H \times W} \) an image of an object, by \( \Omega = \{0, \ldots, H-1\} \times \{0, \ldots, W-1\} \) its domain, and by \( u \in \Omega \) an image pixel. Consider as in [45] a spherical parameterisation of the object surface, where each point on the sphere indexes a different characteristic point of the object, i.e. a landmark. Our goal is to learn a function \( \Phi \) that maps pixels \( u \in \Omega \) to their corresponding landmark indices \( \Phi_u(x) \in S^2 \).

The authors of [45] showed that \( \Phi \) can be learned without manual supervision by requiring it to be invariant with transformations of the image. Namely, consider a random warp \( g : \Omega \to \Omega \) and denote with \( g(x) \) the result of applying the warp to the image.\(^1\) Then, if the map assigns label \( \Phi_u(x) \) to pixel \( u \) of image \( x \), it must assign the same label \( \Phi_{g(u)}(g(x)) \) to pixel \( g(u) \) of the deformed image \( g(x) \). This is because, by construction, pixels \( u \) and \( g(u) \) land on the same object point, and thus contain the same landmark. Hence, we obtain the equivariance constraint \( \Phi_u(x) = \Phi_{g(u)}(g(x)) \).

This version of the equivariance constraint is not quite sufficient to learn meaningful landmarks. In fact, the constraint can be satisfied trivially by mapping all pixels to some fixed point on the sphere. Instead, we must also require landmarks to be distinctive, i.e. identify a unique point in the object. This is captured by the equation:

\[
\forall u, v \in \Omega : \quad v = gu \Leftrightarrow \Phi_u(x) = \Phi_v(g(x)). \tag{1}
\]

**Probabilistic formulation.** For learning, eq. (1) is relaxed probabilistically (fig. 3). Given images \( x \) and \( x' \), define the probability of pixel \( u \) in image \( x \) matching pixel \( v \) in image \( x' \) by normalising the cosine similarity \( \langle \Phi_u(x), \Phi_v(x') \rangle \) of the corresponding landmark vectors:

\[
p(v|u; \Phi, x, x') = \frac{e^{\langle \Phi_u(x), \Phi_v(x') \rangle}}{\int_\Omega e^{\langle \Phi_u(x), \Phi_v(x') \rangle} \, dt}. \tag{2}
\]

\(^1\text{i.e. } (gx)_u = x_{g^{-1}u}.\)
Given a warp \( g \), and image \( x \) and its deformation \( x' = gx \), constraint eq. (1) is captured by the loss:

\[
\mathcal{L}(\Phi; x, x', g) = \frac{1}{|\Omega|^2} \int_{\Omega} \int_{\Omega} \|v - gu\| p(v|u; \Phi, x, x') \, du \, dv
\]  

(3)

where \( \|v - gu\| \) is a distance between pixels. In order to understand this loss, note that \( \mathcal{L}(\Phi; x, x', g) = 0 \) if, and only if, for each pixel \( u \in \Omega \), the probability \( p(v|u; \Phi, x, x') \) puts all its mass on the corresponding pixel \( gu \). Thus minimising this loss encourages \( p(v|u; \Phi, x, x') \) to establish correct deterministic correspondences.

Note that the spread of probability (2) only depends on the angle between landmark vectors. In order to allow the model to modulate this spread directly, the range of function \( \Phi \) is relaxed to be \( \mathbb{R}^3 \). In this manner, estimating longer landmark vectors causes (2) to become more concentrated, and this allows the model to express the confidence of detecting a particular landmark at a certain image location.

**Siamese learning with random warps.** We now explain how (3) can be used to learn the landmark detector function \( \Phi \) given only an unlabelled collection \( X = \{x_1, \ldots, x_n\} \) of images of the object. The idea is to synthesise for each image a corresponding random warp from a distribution \( \mathcal{G} \). Denote with \( \mathcal{P} \) the empirical distribution over the training images; then this amounts to optimising the energy

\[
E(\Phi) = \mathbb{E}_{x \sim \mathcal{P}, g \sim \mathcal{G}} [\mathcal{L}(\Phi; x, gx, g)].
\]  

(4)

Implemented as a neural network, this is a Siamese learning formulation because the network \( \Phi \) is evaluated on both \( x \) and \( gx \).

---

**3.2. From landmarks to descriptors**

Equation (1) says that landmark vectors must be invariant to image transformations and distinctive. Remarkably, exactly the same criterion is often used to define and learn local invariant feature descriptors instead [1]. In fact, if we relax the function \( \Phi \) to produce embeddings in some high-dimensional vector space \( \mathbb{R}^C \), then the formulation above can be used out-of-the-box to learn descriptors instead of landmarks.

Thus the only difference is that landmarks are constrained to be tiny vectors (just points on the sphere), whereas descriptors are usually much higher-dimensional. As argued in section 1, the low dimensionality of the landmark vectors forgets instance-specific details and promotes intra-class generalisation of these descriptors.

The opposite is also true: we can start from any descriptor and turn it into a landmark detector by promoting intra-class generalisation. Using a low-dimensional embedding space is a way to do so, but not the only one, nor the most direct. We propose in the next section an alternative approach.

**3.3. Vector exchangeability**

We now propose our method, Descriptor Vector Exchange, to learn embedding vectors that are distinctive, transformation invariant, and insensitive to intra-class variations, and thus identify object landmarks. The idea is to encourage the sets of embedding vectors extracted from an image to be exchangeable with the ones extracted from another while retaining matching accuracy.

In more detail, let \( (x, x', g) \) be a warped image pair (hence \( x' = gx \)). Furthermore, let \( x_\alpha \) be an auxiliary image, containing an object of the same category as the pair \( (x, x') \), but possibly a different instance. If the embed-
cause some of the pixels in the source image $x$ not to have corresponding matches in the auxiliary image $x_a$.

Next, we integrate this idea in the probabilistic learning formulation given above (fig. 3). We start by matching pixels in the source image $x$ to pixels in the auxiliary image $x_a$ by using the probability $p(w|x; \Phi, x, x_a)$ computed according to eq. (2). Then, we reconstruct the source embedding $\Phi_u(x)$ as the weighted average of the embeddings $\Phi_w(x_a)$ in the auxiliary image, as follows:

$$\hat{\Phi}_u(x|x_a) = \int \Phi_w(x_a)p(w|x; \Phi, x, x_a) dw. \quad (5)$$

Once $\hat{\Phi}_u$ is computed, we use it to establish correspondences between $x$ and $x'$, using eq. (2). This results in the matching probability:

$$p(v|u; \Phi_1, x, x', x_a) = \frac{e^{\langle \Phi_u(x|x_a), \Phi_v(x') \rangle}}{\int_{\mathcal{G}} e^{\langle \Phi_u(x|x_a), \Phi_y(x') \rangle} dt}, \quad (6)$$

This matching probability can be used in the same loss function (3) as before, with the only difference that now each sample depends on $x, x'$ as well as the auxiliary image $x_a$.

**Discussion.** While this may seem a round-about way of learning correspondences, it has two key benefits: as eq. (3) encourages vectors to be invariant and distinctive; in addition to eq. (3), DVE also requires vectors to be compatible between different object instances. In fact, without such a compatibility, the reconstruction (5) would result in a distorted, unmatchable embedding vector. Note that the original formulation of [45] lacks the ability to enforce this compatibility directly.

### 3.4. Using multiple auxiliary images

A potential issue with eq. (6) is that, while image $x'$ can be obtained from $x$ by a synthetic warp so that all pixels can be matched, image $x_a$ is only weakly related to the two. For example, partial occlusions or out of plane rotations may cause some of the pixels in $x$ to not have corresponding pixels in $x_a$.

In order to overcome this issue, we take inspiration from the recent method of [59] and consider not one, but a small set $\{x_a : \alpha \in A\}$ of auxiliary images. Then, the summation in eq. (5) is extended not just over spatial locations, but also over images in this set. The intuition for this approach is that as long as at least one image in the auxiliary image set matches $x$ sufficiently well, then the reconstruction will be reliable.

### 4. Experiments

Using datasets of human faces (section 4.1), animal faces (section 4.3) and a toy robotic arm (section 4.4), we demonstrate the effectiveness of the proposed Descriptor Vector Exchange technique in two ways. First, we show that the learned embeddings work as visual descriptors, matching reliably different views of an object instance. Second, we show that they also identify a dense family of object landmarks, valid not for one, but for all object instances in the same category. Note that, while the first property is in common with traditional and learned descriptors in the spirit of SIFT, the second clearly sets DVE embeddings apart from these.

**Implementation details.** In order to allow for a comparison with the literature, we perform experiments with the deep neural network architecture of [45] (which we refer to as SmallNet). Inspired by the success of the Hourglass model in [54], we also experiment with a more powerful hourglass design (we use the “Stacked Hourglass” design of [32] with a single stack). The weights of both models are learned from scratch using the Adam optimiser [19] for 100 epochs with an initial learning rate of 0.001 and without weight decay. Further details of the architectures are provided in the supplementary material.

#### 4.1. Human faces

First, we consider two standard benchmark datasets of human faces: CelebA [25] and MAFL [56], which is a subset of the former. The CelebA [25] dataset contains over 200k faces of celebrities; we use the former for training and evaluation and exclude embedding quality on the smaller MAFL [56] dataset (19,000 train images, 1,000 test images). Annotations are provided for the eyes, nose and mouth corners. For training, we follow the same procedure used by [45] and exclude any image in the CelebA training set that is also contained in the MAFL test set. Note that we use MAFL annotations only for evaluation and never for training of the embedding function.

We use formulation (6) to learn a dense embedding function $\Phi$ mapping an image $x$ to $C$-dimensional pixel embeddings, as explained above. Note that loss (3) requires sampling transformations $g \in \mathcal{G}$; in order to allow a direct comparison with [45], we use the same random Thin Plate Spline (TPS) warps as they use, obtaining warped pairs $(x, x' = gx)$. We also sample at random one or more auxiliary images $x_a$ from the training set in order to implement DVE.

We consider several cases; in the first, we set $C = 3$ and sample no auxiliary images, using formulation (2), which is the same as [45]. In the second case, we set $C = 16, 32, 64 \gg 3$ but still do not use DVE; in the last case, we use $C = 3, 16, 32, 64$ and also use DVE.
Table 1: Pixel error when matching annotated landmarks across 1000 pairs of images from CelebA (MAFL test set).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1.33</td>
<td>2.89</td>
</tr>
<tr>
<td>16</td>
<td>1.25</td>
<td>5.65</td>
</tr>
<tr>
<td>32</td>
<td>1.26</td>
<td>5.81</td>
</tr>
<tr>
<td>64</td>
<td>1.25</td>
<td>5.68</td>
</tr>
</tbody>
</table>

Qualitative results. In fig. 4 we compute 64D embeddings with SmallNet models trained with or without DVE on AFLW_M images, visualising as in fig. 1 (left). With DVE, matches are accurate despite large intra-class variations. Without DVE, embedding quality degrades significantly. This shows that, by having a category-wide validity, embeddings learned with DVE identify object landmarks rather than mere visual descriptors of local appearance.

![Without DVE - unstable](image1) ![With DVE - maintains robustness](image2)

Figure 4: Learning 64D descriptors without/with DVE

Matching results. Next, we explore the ability of the embeddings learned with SmallNet to match face images. We sample pairs of different identities using MAFL test (1000 pairs total) and consider two cases: First, we match images x, x' of the same identity; since multiple images of the same identity are not provided, we generate them with warps as before, so that the ground-truth correspondence field g is known. We extract embeddings at the annotated keypoint positions from x and match them to their closest neighbour embedding in image x' (searching all pixels in the target). Second, we match images of different identities, again using the annotations. In both cases, we report the mean pixel matching error from the ground truth.

Exhancing the results in table 1 we note several facts. When matching the same identities, higher dimensional embeddings work better than lower (i.e. 3D), including in particular [45]. This is expected as high dimensional embeddings more easily capture instance-specific details; also as expected, DVE does not change the results much as here there are no intra-class variations. When matching different identities, high-dimensional embeddings are rather poor: these descriptors are too sensitive to instance-specific details and cannot bridge intra-class variations correctly. This justifies the choice of a low dimensional embedding in [45] as the latter clearly generalises better across instances. However, once DVE is applied, the performance of the high-dimensional embeddings is much improved, and is in fact better than the low-dimensional descriptors even for intra-class matching [45].

Table 2: Landmark detection results on the MAFL, 300W and AFLW (AFLW_M and AFLW_R splits—see section 4.1 for details). The results are reported as percentage of interocular distance. * report a more conservative evaluation metric (see [6]); † and ‡ use different training data: VoxCeleb [30] and VoxCeleb+ (the union of VoxCeleb and VoxCeleb2 [2]) respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>Unsup. MAFL</th>
<th>AFLW_M</th>
<th>AFLW_R</th>
<th>300W</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCDCN [56]</td>
<td>×</td>
<td>7.95</td>
<td>7.65</td>
<td>5.54</td>
</tr>
<tr>
<td>RAR [30]</td>
<td>×</td>
<td>7.23</td>
<td>7.09</td>
<td>4.94</td>
</tr>
<tr>
<td>MTCNN [55, 54]</td>
<td>×</td>
<td>5.39</td>
<td>6.90</td>
<td>6.31</td>
</tr>
<tr>
<td>Sparse [46]</td>
<td>✓</td>
<td>6.67</td>
<td>10.53</td>
<td>7.97</td>
</tr>
<tr>
<td>Structural Repr. [54]</td>
<td>✓</td>
<td>3.15</td>
<td>6.58</td>
<td></td>
</tr>
<tr>
<td>FAb-Net [49]</td>
<td>†</td>
<td>✓</td>
<td>3.44</td>
<td>-</td>
</tr>
<tr>
<td>Def. AE [42]</td>
<td>✓</td>
<td>5.45</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cond. ImGen. [13]</td>
<td>✓</td>
<td>2.54</td>
<td>6.31</td>
<td></td>
</tr>
<tr>
<td>UDIIT [14]</td>
<td>‡</td>
<td>✓</td>
<td>-</td>
<td>5.37</td>
</tr>
<tr>
<td>Dense 3D [45]</td>
<td>✓</td>
<td>4.02</td>
<td>10.99</td>
<td>8.23</td>
</tr>
<tr>
<td>DVE SmallNet-64D</td>
<td>✓</td>
<td>3.42</td>
<td>8.60</td>
<td>7.79</td>
</tr>
<tr>
<td>DVE Hourglass-64D</td>
<td>✓</td>
<td>2.86</td>
<td>7.53</td>
<td>6.54</td>
</tr>
</tbody>
</table>

Landmark regression. Next, as in [45] and other recent papers, we assess quantitatively how well our embeddings correspond to manually-annotated landmarks in faces. For this, we follow the approach of [45] and add on top of our embedding 50 filters of dimension 1 × 1 × C, converting them into the heatmaps of 50 intermediate virtual points; these heatmaps are in turn converted using a softmax layer to 2C x-y pairs which are finally fed to a linear regressor to estimate manually annotated landmarks. The parameters of the intermediate points and linear regressor are learned using a certain number of manual annotations, but the signal is not back-propagated further so the embeddings remain fully unsupervised.

In detail, after pretraining both the SmallNet and Hourglass networks on the CelebA dataset in an unsupervised
manner, we freeze its parameters and only learn the regressors for MAFL [56]. We then follow the same methodology for the 68-landmark 300-W dataset [40], with 3148 training and 689 testing images. We also evaluate on the challenging AFLW [20] dataset, under the 5 landmark setting. Two slightly different evaluation splits for have been used in prior work: one is the train/test partition of AFLW used in the works of [46], [45] which used the existing crops from MTFL [55] and provides 2,995 faces for testing and 10,122 AFLW faces for training (we refer to this split as AFLW$_M$). The second is a set of re-cropped faces released by [54], which comprises 2991 test faces with 10,122 train faces (we refer to this split as AFLW$_R$). For both AFLW partitions, and similarly to [45], after training for on CelebA we continue with unsupervised pretraining on 10,122 training images from AFLW for 50 epochs (we provide an ablation study to assess the effect of this choice in section 4.2). We report the errors in percentage of inter-ocular distance in table 2 and compare our results to state-of-the-art supervised and unsupervised methods, following the protocol and data selection used in [45] to allow for a direct comparison.

We first see that the proposed DVE method outperforms the prior work that either learns sparse landmarks [46] or 3D dense feature descriptors [45], which is consistent with the results in table 1. Encouragingly, we also see that our method is competitive with the state-of-the-art unsupervised learning techniques across the different benchmarks, indicating that our unsupervised formulation can learn useful information for this task.

4.2. Ablations

In addition to the study evaluating DVE presented in table 1, we conduct two additional experiments to investigate: (i) The sensitivity of the landmark regressor to a reduction in training annotations; (ii) the influence of additional unsupervised pretraining on a target dataset.

**Limited annotation:** We evaluate how many image annotations our method requires to learn landmark localisation in the AFLW dataset, comparing to Dense3D [45] (which shares the SmallNet backbone architecture). To do so, we vary the number of training images across the following range: 1, 5, 10, 20 and up to the whole training set (10,122 in total) and report the errors for each setting in fig. 5. For reference, we also include the supervised CNN baseline from [46] (suppl. material), which consists of a slightly modified SmallNet (denoted SmallNet+ in fig. 5) to make it better suited for landmark regression. Where available, we report the mean and std. deviation over three randomly seeded runs. Further details of this experiment and the SmallNet+ architecture are provided in the suppl. material. While there is considerable variance for very small numbers of annotations, the results indicate that DVE can produce effective landmark detectors with few manual annotations.

**Unsupervised finetuning:** Next we assess the influence of using unsupervised finetuning of the embeddings on a given target dataset, immediately prior to learning to regress landmarks. To do so, we report the performance of several models with and without finetuning on both the AFLW$_M$ and 300W benchmarks in table 3. We see that for AFLW$_M$, this approach (which can be achieved “for free” i.e. without collecting additional annotations) brings a boost in performance. However, it is less effective for 300W, particularly at higher dimensions, having no influence on the performance of the stronger hourglass model.

### Table 3: The effect of unsupervised finetuning on landmark regression performance (errors reported as percentage of inter-ocular distance). Each table entry describes performance without/with finetuning. All methods use DVE.

<table>
<thead>
<tr>
<th>Backbone</th>
<th>Embed. dim</th>
<th>AFLW$_M$</th>
<th>300W</th>
</tr>
</thead>
<tbody>
<tr>
<td>SmallNet</td>
<td>3</td>
<td>11.82 / 11.12</td>
<td>7.66 / 7.20</td>
</tr>
<tr>
<td>SmallNet</td>
<td>16</td>
<td>10.22 / 9.15</td>
<td>6.29 / 5.90</td>
</tr>
<tr>
<td>SmallNet</td>
<td>32</td>
<td>9.80 / 9.17</td>
<td>6.13 / 5.75</td>
</tr>
<tr>
<td>SmallNet+</td>
<td>64</td>
<td>9.28 / 8.60</td>
<td>5.75 / 5.58</td>
</tr>
<tr>
<td>Hourglass</td>
<td>64</td>
<td>8.15 / 7.53</td>
<td>4.65 / 4.65</td>
</tr>
</tbody>
</table>

4.3. Animal faces

To investigate the generalisation capabilities of our method, we consider learning landmarks in an unsupervised manner not just for humans, but for animal faces. To do this, we simply extend the set $X$ of example image to contain images of animals as well.

In more detail, we consider the Animal Faces dataset [43] with images of 20 animal classes and about...
100 images per class. We exclude birds and elephants since these images have a significantly different appearance on average (birds profile, elephants include whole body). We then add additional 8609 additional cat faces from [53], 3506 cat and dog faces from [35], and 160k human faces from CelebA (but keep roughly the same distribution of animal classes per batch as the original dataset). We train SmallNet descriptors using DVE on this data. Here we also found it necessary to use the grouped attention mechanism (section 3.4) which relaxes DVE to project embeddings on a set of auxiliary images rather than just one. In order to do so, we include 16 pairs of images $([x, x'])$ in each batch and we randomly choose a set of 5 auxiliary images for each pair from a separate pool of 16 images. Note that these images have also undergone synthetic warps. Results matching human and cat landmarks to other animals are shown in fig. 6. DVE achieves localisation of semantically-analogous parts across species, with excellent results particularly for the eyes and general facial region.

### 4.4. Roboarm

![Figure 7](image)

Figure 7: An example of descriptor matching on a pair from the roboarm dataset, using blob centres in the first image to locate them in a second image. We show 3D/20D descriptors (columns 2/3) learned with the loss from [45]. The high error of the 20D case is corrected by DVE (last column).

Lastly, we experimented on the animated robotic arm dataset (fig. 7) introduced in [45] to demonstrate the applicability of the approach to diverse data. This dataset contains around 24k images of resolution $90 \times 90$ with ground truth optical flow between frames for training. We use the same matching evaluation of section 4.1 using the centre of the robot’s segments as keypoints for assessing correspondences. We compare models using 3D and 20D embeddings using the formulation of [45] with and without DVE, and finally removing transformation equivariance from the latter (by setting $g = 1$ in eq. (6)).

In this case there are no intra-class variations, but the high-degree of articulation makes matching non-trivial.

Without DVE, 20D descriptors are poor (10.34 error) whereas 3D are able to generalise (1.42). With DVE, however, the 20D descriptors (at 1.25 error) outperform the 3D ones (1.41). Interestingly, DVE is effective enough that even removing transformations altogether (by learning from pairs of identical images using $g = 1$) still results in good performance (1.42) – this is possible because matches must hop through the auxiliary image set $x_\alpha$ which contains different frames.

### 5. Conclusions

We presented a new method that can learn landmark points in an unsupervised way. We formulated this problem in terms of finding correspondences between objects from the same or similar categories. Our method bridges the gap between two seemingly independent concepts: landmarks and local image descriptors. We showed that relatively high dimensional embeddings can be used to simultaneously match and align points by capturing instance-specific similarities as well as more abstract correspondences. We also applied this method to predict facial landmarks in standard computer vision benchmarks as well as to find correspondences across different animal species.

**Acknowledgements.** We thank Almut Sophia Koepke for helpful discussions. We are grateful to ERC StG IDIU-638009, EP/R03298X/1 and AWS Machine Learning Research Awards (MLRA) for support.
References


[34] David Novotny, Diane Larlus, and Andrea Vedaldi. Learning 3d object categories by looking around them. In ICCV, pages 5218–5227, 2017. 3


[43] Zhangzhang Si and Song-Chun Zhu. Learning hybrid image templates (HIT) by information projection. PAMI, 2012. 7


[53] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection - How to effectively exploit shape and texture features. In ECCV, 2008. 8


[56] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning Deep Representation for Face Alignment with Auxiliary Attributes. PAMI, 2016. 5, 6, 7
