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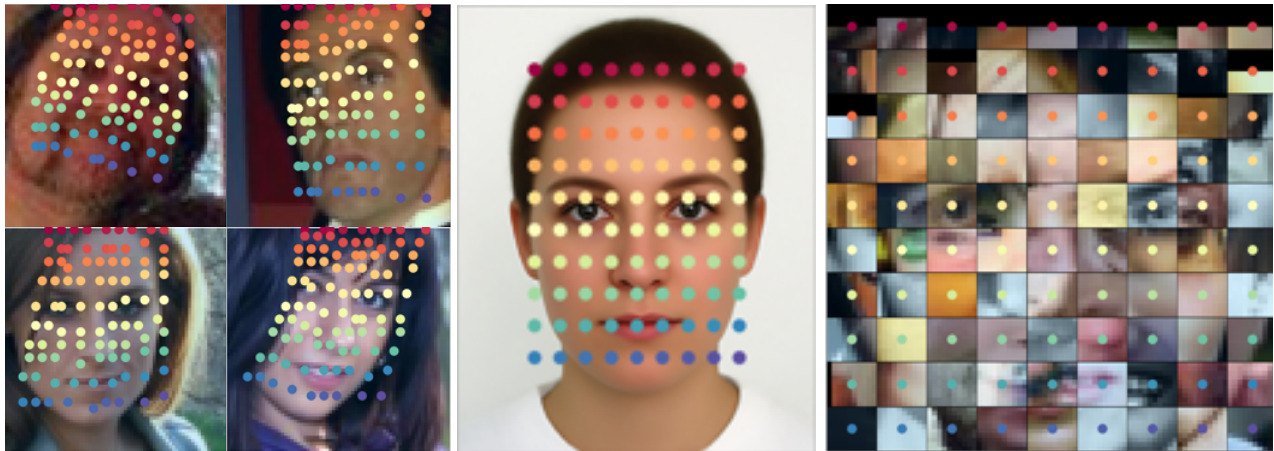VGG, University of Oxford
vedaldi@robots.ox.ac.uk

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 advan-tages 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

---

*Equal Contribution. James was with the VGG during part of this work.

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.
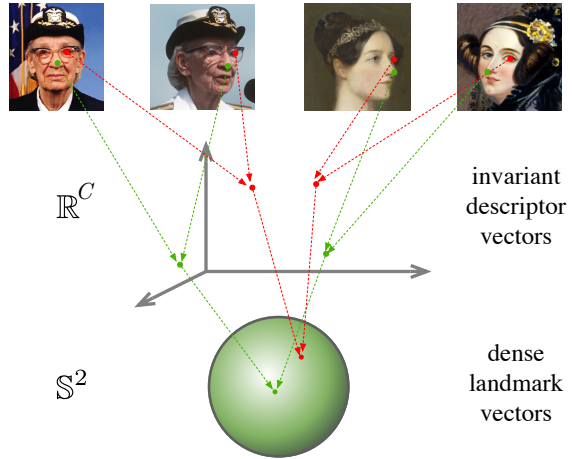

Figure 2: **Descriptor-landmark hierarchy.** A local invariant descriptor maps image pixels to distinctive vectors that are invariant to viewing conditions such as a viewpoint. A dense landmark detector maps pixels to unique points of the object's surface, such as eyes and nose in faces, to points on 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.

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 of a sphere 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 $\mathbf{x} \in \mathbb{R}^{3 \times H \times W}$ an image of an object, by $\Omega = \{0, \dots, H - 1\} \times \{0, \dots, W - 1\}$ 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 Φ that maps pixels $u \in \Omega$ to their corresponding landmark indices $\Phi_u(\mathbf{x}) \in \mathbb{S}^2$.

The authors of [45] showed that Φ 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 $g\mathbf{x}$ the result of applying the warp to the image.[1] Then, if the map assigns label $\Phi_u(\mathbf{x})$ to pixel u of image $\mathbf{x}$, it must assign the same label $\Phi_{gu}(g\mathbf{x})$ to pixel gu of the deformed image $g\mathbf{x}$. This is because, by construction, pixels u and gu land on the same object point, and thus contain the same landmark. Hence, we obtain the *equivariance constraint* $\Phi_u(\mathbf{x}) = \Phi_{gu}(g\mathbf{x})$.

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. to identify a unique point in the object. This is captured by the equation:

$$\forall u, v \in \Omega : \quad v = gu \iff \Phi_u(\mathbf{x}) = \Phi_v(g\mathbf{x}). \quad (1)$$

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

$$p(v|u; \Phi, \mathbf{x}, \mathbf{x}') = \frac{e^{\langle \Phi_u(\mathbf{x}), \Phi_v(\mathbf{x}') \rangle}}{\int_\Omega e^{\langle \Phi_u(\mathbf{x}), \Phi_t(\mathbf{x}') \rangle} dt}. \quad (2)$$

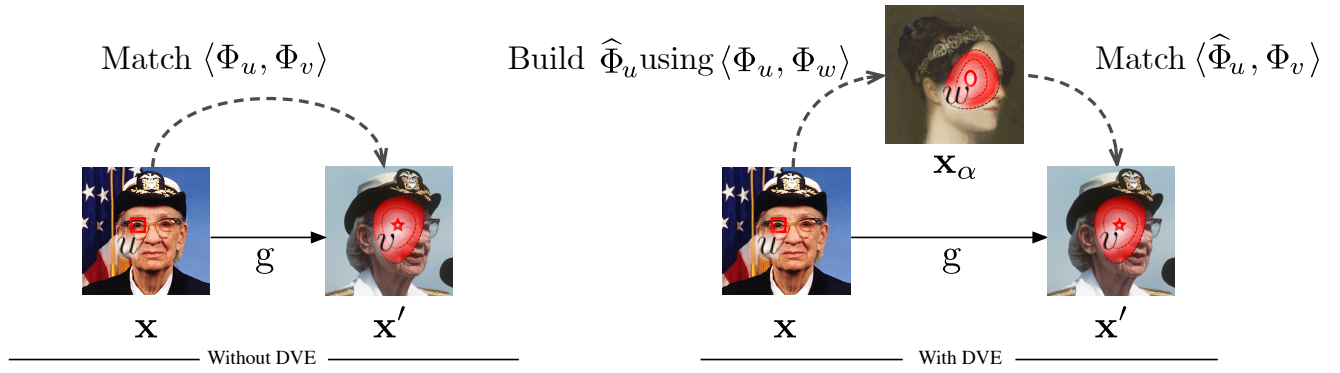[1] I.e. $(g\mathbf{x})_u = \mathbf{x}_{g^{-1}u}$.

Figure 3: We learn a dense embedding $\Phi_u(\mathbf{x}) \in \mathbb{R}^C$ of image pixels. The embedding is learned from pairs of images $(\mathbf{x}, \mathbf{x}')$ related by a known warp $v = g(u)$. Note that in practice, we do not have access to pairs of pairs of images with a known correspondence—thus, throughout this work the warps are generated synthetically. Left: the approach of [45] directly matches embedding $\Phi_u(\mathbf{x})$ from the left image to embeddings $\Phi_v(\mathbf{x}')$ in the right image. Right: DVE replaces $\Phi_u(\mathbf{x})$ from its reconstruction $\widehat{\Phi}_u(\mathbf{x}|\mathbf{x}_\alpha)$ obtained from the embeddings in a third auxiliary image $\mathbf{x}_\alpha$. Importantly, the correspondence with $\mathbf{x}_\alpha$ does not need to be known.

Given a warp g, and image $\mathbf{x}$ and its deformation $\mathbf{x}' = g\mathbf{x}$, constraint eq. (1) is captured by the loss:

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

(3)

where $\|v - gu\|$ is a distance between pixels. In order to understand this loss, note that $\mathcal{L}(\Phi; \mathbf{x}, \mathbf{x}', g) = 0$ if, and only if, for each pixel $u \in \Omega$, the probability $p(v|u; \Phi, \mathbf{x}, \mathbf{x}')$ puts all its mass on the corresponding pixel gu. Thus minimising this loss encourages $p(v|u; \Phi, \mathbf{x}, \mathbf{x}')$ 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 Φ 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.[2]

Siamese learning with random warps. We now explain how (3) can be used to learn the landmark detector function Φ given only an unlabelled collection $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ 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

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

(4)

Implemented as a neural network, this is a Siamese learning formulation because the network Φ is evaluated on both $\mathbf{x}$ and $g\mathbf{x}$.

---

[2]The landmark identity is recovered by normalising the vectors to unit length.

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 Φ 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 $(\mathbf{x}, \mathbf{x}', g)$ be a warped image pair (hence $\mathbf{x}' = g\mathbf{x}$). Furthermore, let $\mathbf{x}_\alpha$ be an *auxiliary image*, containing an object of the same category as the pair $(\mathbf{x}, \mathbf{x}')$, but possibly a different instance. If the embed-

ding function $\Phi_u(\mathbf{x})$ is insensitive to intra-class variations, then the set of embedding vectors $\{\Phi_u(\mathbf{x}) : u \in \Omega\}$ and $\{\Phi_u(\mathbf{x}_\alpha) : u \in \Omega\}$ extracted from any two images should be approximately the same. This means that, in loss (3), we can exchange the vectors $\Phi_u(\mathbf{x})$ extracted from image $\mathbf{x}$ with corresponding vectors extracted from the auxiliary image $\mathbf{x}_\alpha$.

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

$$\widehat{\Phi}_u(\mathbf{x}|\mathbf{x}_\alpha) = \int \Phi_w(\mathbf{x}_\alpha)p(w|u; \Phi, \mathbf{x}, \mathbf{x}_\alpha)\, dw. \quad (5)$$

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

$$p(v|u; \Phi, \mathbf{x}, \mathbf{x}', \mathbf{x}_\alpha) = \frac{e^{\langle \widehat{\Phi}_u(\mathbf{x}|\mathbf{x}_\alpha), \Phi_v(\mathbf{x}') \rangle}}{\int_\Omega e^{\langle \widehat{\Phi}_u(\mathbf{x}|\mathbf{x}_\alpha), \Phi_t(\mathbf{x}') \rangle}\, dt}. \quad (6)$$

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

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 $\mathbf{x}'$ can be obtained from $\mathbf{x}$ by a synthetic warp so that all pixels can be matched, image $\mathbf{x}_\alpha$ is only weakly related to the two. For example, partial occlusions or out of plane rotations may cause some of the pixels in $\mathbf{x}$ to not have corresponding pixels in $\mathbf{x}_\alpha$.

In order to overcome this issue, we take inspiration from the recent method of [59] and consider not one, but a small set $\{\mathbf{x}_\alpha : \alpha \in A\}$ 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 $\mathbf{x}$ 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 well 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 evaluate embedding quality on the smaller MAFL [56] (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 Φ mapping an image $\mathbf{x}$ 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 $(\mathbf{x}, \mathbf{x}' = g\mathbf{x})$. We also sample at random one or more auxiliary images $\mathbf{x}_\alpha$ 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.

Embedding	Same identity		Different identity	
dimension	[45]	+ DVE	[45]	+ DVE
3	1.33	1.36	2.89	3.03
16	1.25	1.28	5.65	2.79
32	1.26	1.29	5.81	2.79
64	1.25	1.28	5.68	2.77

Table 1: Pixel error when matching annotated landmarks across 1000 pairs of images from CelebA (MAFL test set).

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.

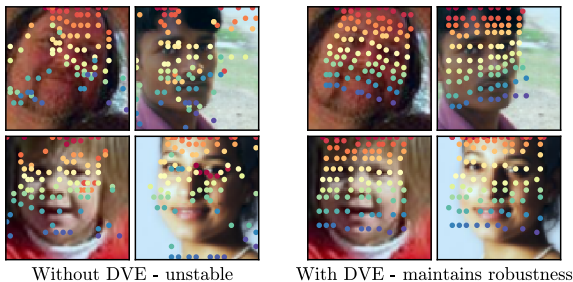Without DVE - unstable   With DVE - maintains robustness

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 $\mathbf{x}, \mathbf{x}'$ 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 $\mathbf{x}$ and match them to their closest neighbour embedding in image $\mathbf{x}'$ (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.

Examining 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

Method	Unsup.	MAFL	$AFLW_M$	$AFLW_R$	300W
TCDCN [56]	×	7.95	7.65	–	5.54
RAR [50]	×		7.23	–	4.94
MTCNN [55, 54]	×	5.39	6.90	–	–
Wing Loss [6]*	×	-	-	-	4.04
Sparse [46]	✓	6.67	10.53	–	7.97
Structural Repr. [54]	✓	3.15	–	6.58	–
FAb-Net [49]‡	✓	3.44	–	–	5.71
Def. AE [42]	✓	5.45	–	–	–
Cond. ImGen. [13]	✓	**2.54**	–	6.31	–
UDIT [14]†	✓	-	-	-	5.37
Dense 3D [45]	✓	4.02	10.99	10.14	8.23
DVE SmallNet-64D	✓	3.42	8.60	7.79	5.75
DVE Hourglass-64D	✓	2.86	**7.53**	6.54	**4.65**

Table 2: Landmark detection results on the MAFL, 300W and AFLW ($AFLW_M$ and $ALFW_R$ 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.

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].

Overall, the embeddings learned with DVE have both better intra-class and intra-instance matching performance than [45], validating our hypothesis and demonstrating that our method for regularising the embedding is preferable to simply constraining the embedding dimensionality.

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 $1 \times 1 \times C$, converting them into the heatmaps of 50 intermediate virtual points; these heatmaps are in turn converted using a softargmax 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 a 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
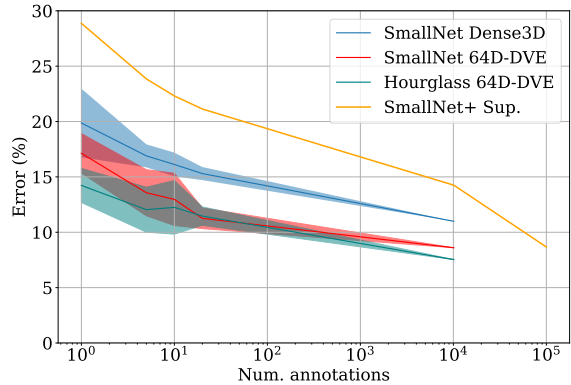

Figure 5: The effect of varying the number of annotated images used for different methods on $AFLW_M$, incorporating the Supervised CNN baseline from [46] (suppl. material).

Backbone	Embed. dim	$AFLW_M$	300W
SmallNet	3	11.82 / 11.12	7.66 / 7.20
SmallNet	16	10.22 / 9.15	6.29 / 5.90
SmallNet	32	9.80 / 9.17	6.13 / 5.75
SmallNet	64	9.28 / 8.60	5.75 / 5.58
Hourglass	64	8.15 /7.53	4.65 / 4.65

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.

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.

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 $\mathcal{X}$ 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

Figure 6: Top: Five landmarks are manually annotated in the top-left image (human) and matched using our unsupervised embedding to a number of animals. Bottom: same process, but using a cat image (bottom left) as query.

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 $(\mathbf{x}, \mathbf{x}')$ 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.
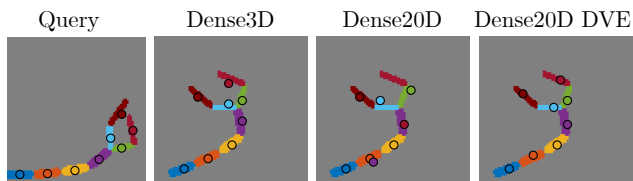
4.4. Roboarm

Figure 7: 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 90×90 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

Dimensionality	[45]	+ DVE	- transformations
3	1.42	1.41	1.69
20	10.34	1.25	1.42

Table 4: Results on Roboarm, including an experiment ignoring optical flow (right).

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 $\mathbf{x}_\alpha$ 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.

References

[1] Matthew Brown, Gang Hua, and Simon Winder. Discriminative learning of local image descriptors. *PAMI*, 2010. 4

[2] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 6

[3] Timothy F. Cootes, Christopher J. Taylor, David H. Cooper, and Jim Graham. Active shape models: their training and application. *CVIU*, 1995. 3

[4] Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, 2005. 3

[5] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object Detection with Discriminatively Trained Part Based Models. *PAMI*, 2010. 3

[6] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *CVPR*, pages 2235–2245, 2018. 6

[7] Rob Fergus, Pietro Perona, and Andrew Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003. 3

[8] Martin Grundl. Average faces. http://www.beautycheck.de/cmsms/index.php/durchschnittsgesichter. [Online; accessed 2019]. 1

[9] Rıza Alp Güler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. In *CVPR*, pages 6799–6808, 2017. 2

[10] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow. In *CVPR*, 2016. 3

[11] Tal Hassner, Viki Mayzels, and Lihi Zelnik-Manor. On sifts and their scales. In *CVPR*, pages 1522–1528. IEEE, 2012. 3

[12] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial Transformer Networks. In *NeurIPS*, 2015. 3

[13] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Advances in Neural Information Processing Systems*, pages 4020–4031, 2018. 3, 6

[14] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Learning landmarks from unaligned data using image translation. *arXiv preprint arXiv:1907.02055*, 2019. 6

[15] Simon Jenni and Paolo Favaro. Self-supervised feature learning by learning to spot artifacts. In *CVPR*, pages 2733–2742, 2018. 1

[16] Angjoo Kanazawa, David W. Jacobs, and Manmohan Chandraker. WarpNet: Weakly supervised matching for single-view reconstruction. In *CVPR*, 2016. 3

[17] Asako Kanezaki, Yasuyuki Matsushita, and Yoshifumi Nishida. Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In *CVPR*, pages 5010–5019, 2018. 1

[18] Ira Kemelmacher-Shlizerman and Steven M. Seitz. Collection flow. In *CVPR*, 2012. 3

[19] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[20] Martin Koestinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Annotated facial landmarks in the wild. In *ICCV workshops*, 2011. 7

[21] Erik G Learned-Miller. Data driven image models through continuous joint alignment. *PAMI*, 2006. 3

[22] Bastian Leibe, Ales Leonardis, and Bernt Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV Workshops*, 2004. 3

[23] Karel Lenc and Andrea Vedaldi. Learning covariant feature detectors. In *ECCV Workshop on Geometry Meets Deep Learning*, 2016. 2

[24] Ce Liu, Jenny Yuen, and Antonio Torralba. SIFT Flow: Dense correspondence across scenes and its applications. *PAMI*, 2011. 3

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5

[26] Jonathan L Long, Ning Zhang, and Trevor Darrell. Do convnets learn correspondence? In *Advances in Neural Information Processing Systems*, pages 1601–1609, 2014. 3

[27] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2, 3

[28] Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, Jiri Matas, Frederik Schaffalitzky, Timor Kadir, and Luc Van Gool. A comparison of affine region detectors. *IJCV*, 65(1-2):43–72, 2005. 2

[29] Hossein Mobahi, Ce Liu, and William T. Freeman. A Compositional Model for Low-Dimensional Image Set Representation. *CVPR*, 2014. 3

[30] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017. 6

[31] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, 2015. 3

[32] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 5, 13

[33] David Novotny, Samuel Albanie, Diane Larlus, and Andrea Vedaldi. Self-supervised learning of geometrically stable features through probabilistic introspection. In *CVPR*, pages 3637–3645, 2018. 3

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

[35] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 8

[36] Mattis Paulin, Matthijs Douze, Zaid Harchaoui, Julien Mairal, Florent Perronin, and Cordelia Schmid. Local convolutional features with unsupervised training for image retrieval. In *ICCV*, pages 91–99, 2015. 2

[37] Yigang Peng, Arvind Ganesh, John Wright, Wenli Xu, and Yi Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *PAMI*, 34(11), 2012. 3

[38] Philip Pritchett and Andrew Zisserman. Wide baseline stereo matching. In *ICCV*, pages 754–760. IEEE, 1998. 2

[39] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *CVPR*, pages 6148–6157, 2017. 3

[40] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *CVPR-W*, 2013. 7

[41] Tanner Schmidt, Richard Newcombe, and Dieter Fox. Self-supervised visual descriptor learning for dense correspondence. *IEEE Robotics and Automation Letters*, 2(2):420–427, 2017. 3

[42] Zhixin Shu, Mihir Sahasrabudhe, Rıza Alp Güler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *ECCV*, pages 650–665, 2018. 2, 3, 6

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

[44] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *ECCV*, pages 438–451. Springer, 2010. 3

[45] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *Advances in Neural Information Processing Systems*, pages 844–855, 2017. 2, 3, 4, 5, 6, 7, 8, 12, 13

[46] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, pages 5916–5925, 2017. 1, 3, 6, 7, 13

[47] Engin Tola, Vincent Lepetit, and Pascal Fua. A fast local descriptor for dense matching. In *CVPR*, pages 1–8. IEEE, 2008. 2

[48] Tinne Tuytelaars, Luc Van Gool, Luk D'haene, and Reinhard Koch. Matching of affinely invariant regions for visual servoing. In *ICRA*, volume 2, pages 1601–1606, 1999. 2

[49] O. Wiles, A. S. Koepke, and A. Zisserman. Self-supervised learning of a facial attribute embedding from video. In *Proc. BMVC*, 2018. 3, 6

[50] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Kassim. Robust Facial Landmark Detection via Recurrent Attentive-Refinement Networks. In *ECCV*, 2016. 6

[51] Christopher Zach, Manfred Klopschitz, and Marc Pollefeys. Disambiguating visual relations using loop constraints. In *CVPR*, pages 1426–1433. IEEE, 2010. 3

[52] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *CVPR*, pages 4353–4361. IEEE, 2015. 2

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

[54] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *CVPR*, 2018. 3, 5, 6, 7

[55] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014. 6, 7

[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

[57] Tinghui Zhou, Yong Jae Lee, Stella X Yu, and Alyosha A Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *CVPR*, pages 1191–1200, 2015. 3

[58] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In *CVPR*, pages 117–126, 2016. 3

[59] Bohan Zhuang, Lingqiao Liu, Yao Li, Chunhua Shen, and Ian Reid. Attend in groups: a weakly-supervised deep learning framework for learning from web data. In *CVPR*, pages 1878–1887, 2017. 5

Figure 8: Additional images querying manual annotations on a human and finding the matching descriptors on animal faces, images are selected randomly from the validation set and include some more severe failure cases (eg mouse mouth row 1 col 2, dog eye row 5 col 8).

Appendix

A. Animal Faces

Here we present additional qualitative results on the animal faces dataset in fig. 8. The leftmost column shows the human faces with their annotated landmarks (drawn as coloured circles) which are matched to a set of queried animal faces in the remaining columns. The correspondent matches are depicted in the same colour with the manual landmarks. We observe that our method achieves to find many semantically meaningful matches in spite of wide variation in the appearance across different species.

B. Roboarm details

We showed results for an experiment showing that the use of optical flow (ground truth flow in this dataset) may not be essential. In this setup we set $\mathbf{x} = \mathbf{x}'$ when training, meaning the same image is used rather than two consecutive frames, and the transformation is the identity $g = 1$. However we still explicitly ignore the background region (which is otherwise achieved by ignoring areas of zero flow). Our hope is that DVE, which has the effect of searching for a matching descriptor in a third image $\mathbf{x}_\alpha$, will be able to stand in for explicit matches given by g. The results appear to confirm this. Surprisingly we can even obtain results matching [45] without flow information, albeit using a higher dimensionality 20D descriptor. However the highest performance is still obtained by using the flow in addition to DVE, therefore we use flow (from synthetic warps) in experiments on faces.

C. Limited Annotation experiments

The numbers corresponding to the figure shown in section 4 of the paper portraying the effect of varying the number of annotated images are given in table 5. To allow the experiment to be reproduced, the list of the randomly sampled annotations is provided on the project page http://www.robots.ox.ac.uk/~vgg/research/DVE.

Num. images	Dense 3D [45]	SmallNet 64D DVE	Hourglass 64D DVE	Smallnet+ Sup.
1	19.87 ± 3.10	17.13 ± 1.78	14.23 ± 1.54	28.87
5	16.90 ± 1.04	13.57 ± 2.08	12.04 ± 2.03	32.85
10	16.12 ± 1.07	12.97 ± 2.36	12.25 ± 2.42	22.31
20	15.30 ± 0.59	11.26 ± 0.93	11.46 ± 0.83	21.13
AFLW$_M$ (10,122)	10.99	8.80	7.53	14.25
CelebA + AFLW$_M$ ($>$ 100k)	-	-	-	8.67

Table 5: Error (% inter-ocular distance) Varying the number of images used for training (AFLW$_M$). The errors are reported in the form (mean $\pm$ std.), where the statistics are computed three randomly seeded samples of annotations. The general indication is that most of the information has been encoded in the unsupervised stage.

D. Architecture details

The *SmallNet* architecture, which was used in in [46, 45] consists of a set of seven convolutional layers with 20, 48, 64, 80, 256 before C filters are used to produce a C-dimensional embedding. The second, third and fourth layers are dilated by 2, 4 and 2 resp. Every convolutional layer except the last is followed by batch norm and a ReLU. Following the first convolutional layer, features are downsampled via a 2×2 max pool layer using a stride of 2. Consequently, for an input size of $H \times W \times 3$, the size of the output $\frac{H}{2} \times \frac{W}{2} \times C$.

The *SmallNet+* architecture, introduced in [46], is a slightly modified version of SmallNet, which further includes pooling layers with a stride of 2 after each of the first three convolutional layers, and operates on an input of size 64×64.

The *Houglass* architecture was introduced (in its "stacked" formulation) by [32]. We use this network with a single stack, operating on inputs of size 96×96 and using preactivation residuals.

The code implementing the architectures used in this work can be found via the project page: http://www.robots.ox.ac.uk/~vgg/research/DVE.

D.1. Preprocessing details

For all datasets, the inputs to *SmallNet* are then resized to 100×100 pixels then centre-cropped to 70×70 pixels (this cropping is done after warping during training). The inputs to *Hourglass* are resized to 136×136 pixels then centre-cropped to 96×96 pixels. For the particular case of the CelebA face crops, which contain a good deal of surrounding context with varied backgrounds, faces are additionally preprocessed by removing the top 30 pixels and bottom 10 vertically from the 218×178 image before resizing. For 300-W we make the ground truth bounding box square (setting the height to equal the width) and then add more context on each side such that the original width occupies the central 52% of the resulting image.