Supplementary Material: Unsupervised learning of landmarks by Descriptor Vector Exchange

James Thewlis*
University of Oxford
james@unitary.ai

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

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

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

1. Animal Faces

Here we present additional qualitative results on the animal faces dataset in fig. 1. 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.

Figure 1. 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).

2. 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

*Equal Contribution

Num. images	Dense 3D [2]	SmallNet 64D DVE	Hourglass 64D DVE	Smallnet+ Sup.
1	19.87 $\pm$3.10	17.13 $\pm$1.78	14.23 $\pm$1.54	28.87
5	16.90 $\pm$1.04	13.57 $\pm$2.08	12.04 $\pm$2.03	32.85
10	16.12 $\pm$1.07	12.97 $\pm$2.36	12.25 $\pm$2.42	22.31
20	15.30 $\pm$0.59	11.26 $\pm$0.93	11.46 $\pm$0.83	21.13
AFLW$_M$ (10,122)	10.99	8.80	7.53	14.25
CelebA + AFLW$_M$ ($>$ 100k)	-	-	-	8.67

Table 1. 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.

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 [2] 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.

3. 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 1. 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.

4. Architecture details

The *SmallNet* architecture, which was used in in [3, 2] 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 [3], 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 [1]. 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.

4.1. Preprocessing details

For all datasets, the inputs to *SmallNet* are then resized to 100×100 pixels then centre-cropped to 70×70 pixels. 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.

References

[1] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 2

[2] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *Advances in Neural Information Processing Systems*, pages 844–855, 2017. 2

[3] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, pages 5916–5925, 2017. 2