Supplementary Material: Unsupervised learning of object landmarks by factorized spatial embeddings

James Thewlis
University of Oxford
jdt@robots.ox.ac.uk

Hakan Bilen
University of Oxford
hbilen@robots.ox.ac.uk

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

1. Introduction

In this supplementary material we elaborate on several details regarding the experimental setup, provide an additional comparison with training a supervised network on small numbers of images and present numerous images giving a qualitative look at the performance of our method. It is organized as follows: Sec. 2 gives the additional details and hyperparameters, Sec. 3 compares quantitatively with a supervised network and qualitative results are shown in Sec. 4. We evaluate the learned features on face segmentation in Sec. 5.

2. Experimental details

As described in Section 4.1 of the main text, we generate a pair of warps (g_1, g_2). These are parameterized as Thin Plate Spline warps, which models the deformation of several keypoints along with an affine component. We sample all parameters from a gaussian with zero mean and the given standard deviations unless otherwise stated. The source keypoints are a 10×10 regular grid (5×5 for MNIST), whereas each element of the parameter vector defining the destination keypoints is sampled with standard deviation $\sigma_{g_i,w}$. For each element we then add with 50% probability an additional perturbation sampled with standard deviation $\sigma_{g_i,W}$.

The affine component is parameterised as a similarity transform with rotation standard deviation $\sigma_{g_i,r}$ degrees, translation $\sigma_{g_i,t}$, and scale $\sigma_{g_i,s}$ with mean 1. Note we operate with normalized coordinated in the range $[-1, 1]$. Values are shown in Table 1. For faces and cats the input image dimensions are 100×100, which are then cropped after warping to 80×80. For MNIST the input images are resized to 35×35 then padded with a 5 pixel black border to be 45×45. For shoes the 64×64 initial images are padded with a 15 pixel white border to be 94×94.

The pooling layer prior to the diversity loss has pooling window size 5×5 in all networks except for MNIST and the AFLW 51 landmark network which have 3×3 (resulting in denser coverage of the face area, fig. 8).

	g_i	$\sigma_{g_i,w}$	$\sigma_{g_i,W}$	$\sigma_{g_i,r}$	$\sigma_{g_i,t}$	$\sigma_{g_i,s}$
Faces	g_1	0.001	0.001	0°	0	0
	g_2	0.001	0.01	20°	0.1	0.05
MNIST	g_1	0.005	0.01	15°	0.1	0.05
	g_2	0.005	0.02	20°	0.1	0.05

Table 1. Standard deviations used for sampling warp parameters.

Labelled Images	Sup. Net	Unsup. + Regressor
CelebA + AFLW	8.67	—
AFLW (10,122)	14.25	10.53
20	21.13	13.28
10	22.31	13.85
5	23.85	12.94
1	28.87	14.79

Table 2. Results on AFLW (2995 images, 5 landmarks), varying the number of images used to train both a supervised network from scratch and a regressor on top of our unsupervised landmarks.

3. Supervised Network Comparison

In order to further evaluate the advantage of our unsupervised pre-training when a limited number of labelled images are used for subsequent supervised training, we compare to training a supervised network from scratch on the same images (Table 2 and fig. 1). The results reported in the main text adapted our unsupervised architecture with the addition of a final pooling layer (stride 2) and fully connected layer, achieving 23.85 error for 20 images. Here we train a network more comparable to existing supervised landmark networks by including pooling layers (stride 2) after the first three convolutional layers and taking a 64×64 input. It achieves results comparable to existing work when trained on many images and evaluated on AFLW (8.67, compare to Table 3 in the main paper) and error of 21.13 on 20 images. This confirms the advantage of our approach in the case of limited labelled data.

4. Qualitative Results

We show additional images displaying the results of our method on different datasets and with different numbers of

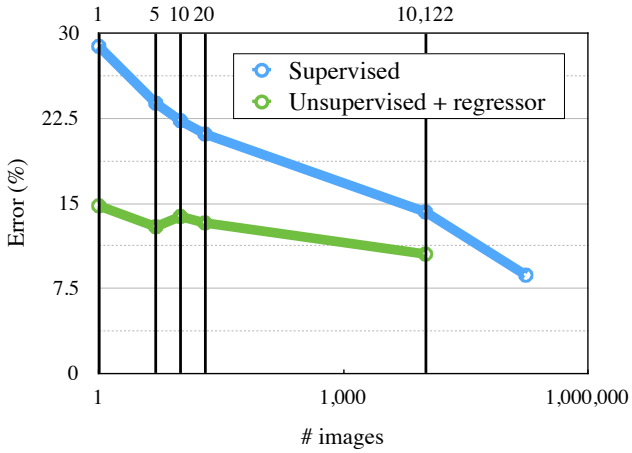

Figure 1. The same data as table 2 in graphical format.

unsupervised landmarks.

The MNIST dataset of handwritten digits provides a simple setting in which to demonstrate the ability of our approach to identify landmarks across variations in writing style. We train separate networks for the digits 3, 5, and 6. The training data is augmented with Thin Plate Spline transformations and similarity transforms (parameters in Table 1). For each digit we use 1000 images for validation and the rest (around 5000) for training. As shown in fig. 3 the discovered landmarks are robust to rotations and significant differences in style.

To complement the examples of a 10-landmark network on cat faces in the main paper, we also show a network with 20 landmarks (fig. 4).

For the CelebA faces dataset (MAFL test subset) we show examples of a 30-landmark network (fig. 5) and the results of training our regressor with varying numbers of landmarks (fig. 7). For the 300-W dataset we show regression examples for a 30-landmark network (fig. 6). We also show the result of the 51-landmark network finetuned on AFLW and the regressor predictions (fig. 8).

In order to evaluate the effectiveness of our network in cases of illumination variation, we apply our 10-landmark CelebA network on frontal faces from the Cropped Extended Yale B[1] dataset. This dataset represents a significantly different domain to that used for training, being grayscale and tightly cropped. Nevertheless, with the more moderate lighting variations we get consistent landmarks. Failure occurs in the cases of hard shadows where there is little resolvable detail in areas of the face. We can fix this failure by finetuning to the target dataset, whereupon landmarks are predicted well across illumination variants. This

[1]Georghiades, A. S., Belhumeur, P. N., & Kriegman, D. J. From few to many: Illumination cone models for face recognition under variable lighting and pose. PAMI 2001
Lee, K. C., Ho, J., & Kriegman, D. J. Acquiring linear subspaces for face recognition under variable lighting. PAMI 2005

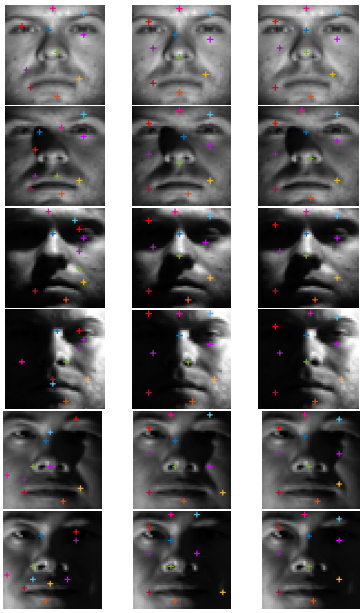

Figure 2. YaleB: Predicted landmarks on two Yale B subjects (held out during finetuning). Column 1: Original CelebA network, with poor results in shadows. Column 2: Finetuning from synthetic warps. Column 3: Finetuning from pairs with different lighting conditions.

	Epoch 5	Epoch 30	Best
From Scratch	78.10%	86.15%	93.59%
Pretrained	90.64%	92.12%	94.46%
Pretrained+ft	90.84%	92.41%	94.82%

Table 3. Pixel accuracy on HELEN when training from scratch, pretraining using our method (Conv 1-3 frozen) and pretraining while finetuning all layers

	20 Images	50 Images
From Scratch	86.52%	86.91%
Pretrained	90.24%	90.63%

Table 4. Pixel accuracy on HELEN segmentation for a limited number of training images.

finetuning can be done using synthetic warps as in the main paper, however we also note that in the cases of datasets like Yale B which offer aligned pairs of the same subject, we can simply train based on the identity transformation between aligned images having different lighting conditions. Both methods give qualitatively good results as shown in fig. 2.

5. Evaluating learned features

We would like to know if the features obtained using our method are useful for other tasks. For this we use the task of face segmentation using the HELEN[2] dataset. We resize the images but do not further crop or preprocess them. We

[2]Smith, B. M., Zhang, L., Brandt, J., Lin, Z., & Yang, J. Exemplar-based face parsing. ICCV 2013

use our pretrained 50-landmark CelebA network with the first three layers frozen and replace the last layer with a 10-way spatial classification. We get 94.46% pixel accuracy, compared to 93.59% for the same network configuration trained from scratch. When we finetune all layers, accuracy increases further to 94.82%. This shows that the initial features learned are useful for general purpose face-based tasks, and that the learned weights are suitable as a starting point for further adaptation. Convergence is also a lot quicker when pretrained as shown in Table 3. An additional advantage of pretraining is that it allows training with fewer images, which we show for 20 and 50 images in Table 4.

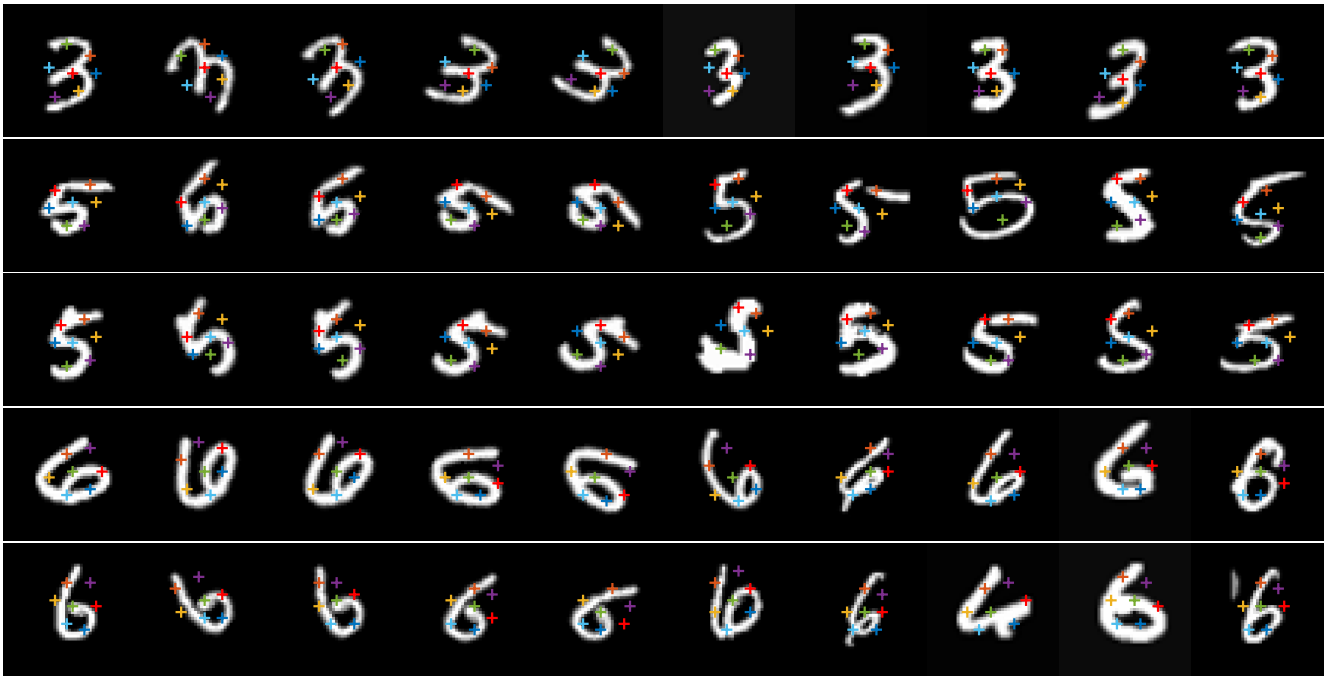

Figure 3. Three 7-landmark networks on MNIST (digits 3,5,6). The first five columns show rotations of the same instance ($0°$, $-50°$, $-30°$, $30°$, $50°$) the rest show arbitrary instances.

Figure 4. 20-landmark cat network

Figure 5. 30-landmark network on CelebA. Row 1: synthetic warps. Rows 2-3: rotations. Rows 4-6: arbitrary instances.

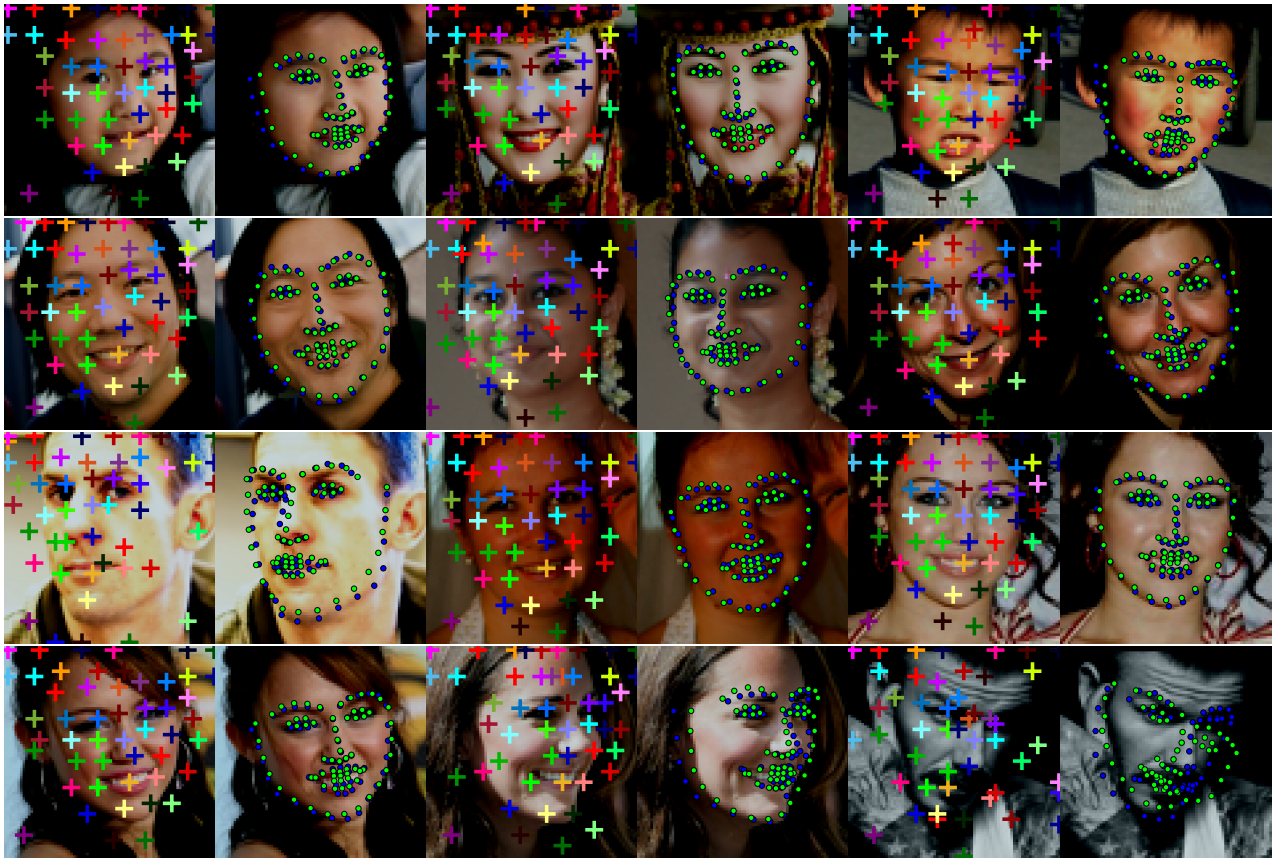

Figure 6. 30-landmark network and regressor output on 300-W. Green circles are predictions, blue circles are ground truth. The last example shows a failure case.

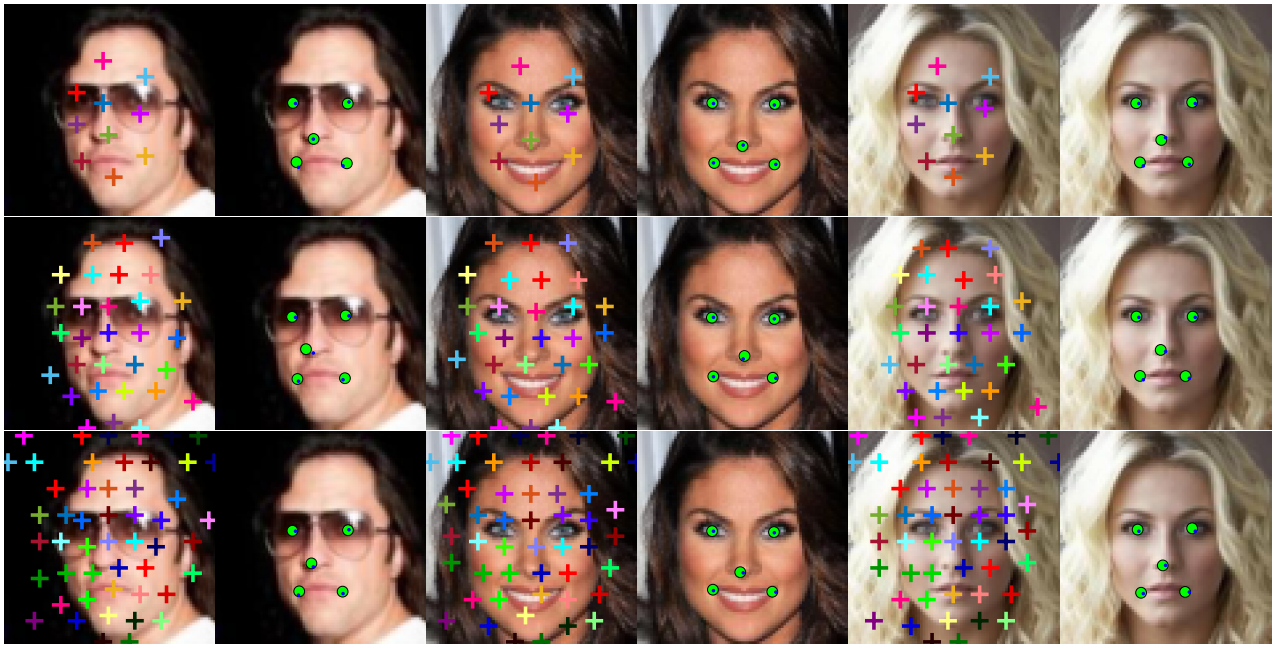

Figure 7. Unsupervised landmarks and regressor predictions for 10, 30 and 50 landmark networks in rows 1, 2 and 3 respectively. Green circles are predictions, blue circles ground truth.

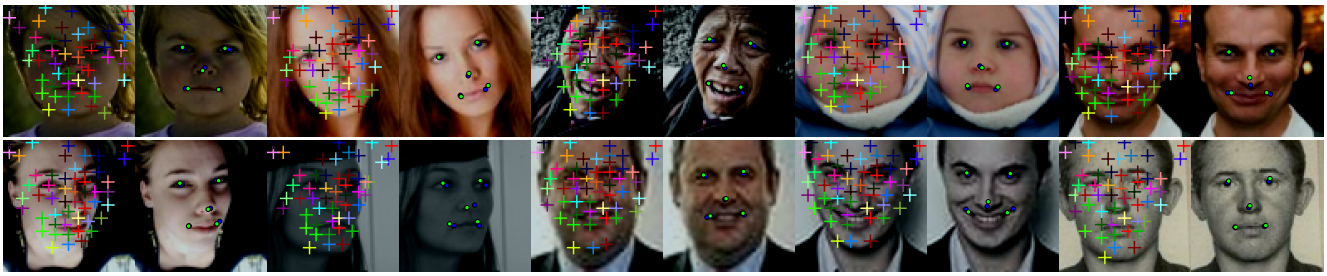

Figure 8. AFLW: Unsupervised landmarks from 51-landmark network and regressor predictions. Green circles are predictions, blue circles ground truth.