Self-supervised Learning of Interpretable Keypoints from Unlabelled Videos

Citation for published version:
https://doi.org/10.1109/CVPR42600.2020.00881

Digital Object Identifier (DOI):
10.1109/CVPR42600.2020.00881

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Self-supervised Learning of Interpretable Keypoints from Unlabelled Videos

Tomas Jakab
Visual Geometry Group
University of Oxford
tomj@robots.ox.ac.uk

Ankush Gupta
DeepMind, London
ankushgupta@google.com

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

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

Abstract

We propose a new method for recognizing the pose of objects from a single image that for learning uses only unlabelled videos and a weak empirical prior on the object poses. Video frames differ primarily in the pose of the objects they contain, so our method distills the pose information by analyzing the differences between frames. The distillation uses a new dual representation of the geometry of objects as a set of 2D keypoints, and as a pictorial representation, i.e., a skeleton image. This has three benefits: (1) it provides a tight ‘geometric bottleneck’ which disentangles pose from appearance, (2) it can leverage powerful image-to-image translation networks to map between photometry and geometry, and (3) it allows to incorporate empirical pose priors in the learning process. The pose priors are obtained from unpaired data, such as from a different dataset or modality such as mocap, such that no annotated image is ever used in learning the pose recognition network. In standard benchmarks for pose recognition for humans and faces, our method achieves state-of-the-art performance among methods that do not require any labelled images for training. Project page: http://www.robots.ox.ac.uk/~vgg/research/unsupervised_pose/

1. Introduction

Learning with limited or no external supervision is one of the most significant open challenges in machine learning. In this paper, we consider the problem of learning the 2D geometry of object categories such as humans and faces using raw videos and as little additional supervision as possible. In particular, given as input a number of videos centred on the object, the goal is to learn automatically a neural network that can predict the pose of the object from a single image.

Learning from unlabelled images requires a suitable supervisory signal. Recently, [25] noted that during a video an object usually maintains its intrinsic appearance but changes its pose. Hence, the concept of pose can be learned by modelling the differences between video frames. They formulate this as conditional image generation. They extract a small amount of information from a given target video frame via a tight bottleneck which retains pose information while discarding appearance. For supervision, they reconstruct the target frame from the extracted pose, similar to an auto-encoder. However, since pose alone does not contain sufficient information to reconstruct the appearance of the object, they also pass to the generator a second video frame from which the appearance can be observed.

In this paper, we also consider a conditional image generation approach, but we introduce a whole new design for the model and for the ‘pose bottleneck’. In particular, we adopt a dual representation of pose as a set of 2D object coordinates, and as a pictorial representation of the 2D coordinates in the form of a skeleton image. We also define a differentiable
skeleton generator to map between the two representations. This design is motivated by the fact that, by encoding pose labels as images we can leverage powerful image-to-image translation networks [77] to map between photometry and geometry. In fact, the two sides of the translation process, namely the input image and its skeleton, are spatially aligned, which is well known to simplify learning by a Convolutional Neural Network (CNN) [77]. At the same time, using 2D coordinates provides a very tight bottleneck that allows the model to efficiently separate pose from appearance.

The pose bottleneck is further controlled via a discriminator, learned adversarially. This has the advantage of injecting prior information about the possible object poses in the learning process. While acquiring this prior may require some supervision, this is separate from the unlabelled videos used to learn the pose recognizer — that is, our method is able to leverage unpaired supervision. In this way, our method outputs poses that are directly interpretable.

2. Related work

We consider pose recognition, intended as the problem of predicting the 2D pose of an object from a single image. Approaches to this problem must be compared in relation to (1) the type of supervision, and (2) which priors they use. There are three broad categories for supervision: full supervision when the training images are annotated with the same labels that one wishes to predict; weak supervision when the predicted labels are richer than the image annotations; and no supervision when there are no image annotations. For the prior, methods can use a prior model learned from any kind of data or supervision, an empirical prior, or no prior at all.

Based on this definition, our method is unsupervised and uses an empirical prior. Next, we relate our work to others, dividing them by the type of supervision used (our method falls in the last category).

Full supervision. Several fully-supervised methods leverage large annotated datasets such as MS COCO Keypoints [31], Human3.6M [23], MPII [2] and LSP [27]. They generally do not use a separate prior as the annotations themselves capture one empirically. Some methods use pictorial structures [12] to model the pose objects [1, 37, 40, 41, 48, 70]. Others use a CNN to directly regress keypoint coordinates [59], keypoint confidence maps [58], or other relations between keypoints [9]. Others again apply networks iteratively to refine heatmaps for single [3, 6, 8, 36, 39, 57, 66] and multi-person settings [7, 22]. Our method does not use any annotated image to learn the pose recognizer.

Weak supervision. A typical weakly-supervised method is the one of Kanazawa et al. [29]: they learn to predict dense 3D human meshes from sparse 2D landmark annotations. They use two priors: SMPL [32] parametric human mesh model, and a prior on 3D poses acquired via adversarial learning from mocap data. Analogous works include [16, 17, 18, 46, 49, 61, 65, 69].

All such methods use a prior trained on unpaired data, as we do. However, they also use additional paired annotations such as 2D keypoints or relative depth relations [46]. Furthermore, in most cases they use a fully-fledged 3D prior such as SMPL human [32] or Basel face [38] models, while we only use an empirical prior in the form of example 2D keypoint configurations.

No supervision. Other methods use no supervision, and some no data-driven prior either. The works of [28, 45, 50, 67] learn to match pairs of images of an object, but they do not learn geometric invariants such as keypoints. [54, 55, 56] do learn sparse and dense landmarks, also without any annotation. The method of [53] does not use image annotations, but uses instead synthetic views of 3D models as prior, which we do not require.

Some of these methods use conditional image generation as we do. Jakab & Gupta et al. [25], the most related, is described in the introduction. Lorenz et al. [33], Zhang et al. [74] develop an auto-encoding formulation to discover landmarks as explicit structural representations for a given
image and use them to reconstruct the original image. Shu et al. [50], Wiles et al. [67] learn a dense deformation field for faces. Our method differs from those in the particular nature of the model and geometric bottleneck; furthermore, due to our use of a prior, we are able to learn out-of-the-box landmarks that are ‘semantically meaningful’; on the contrary, these approaches must rely on at least some paired supervision to translate between the unsupervised and ‘semantic’ landmarks. We also outperform these approaches in landmark detection quality.

Adversarial learning. Our method is also related to adversarial learning, which has proven to be useful in image labelling [14, 20, 21, 62, 63] and generation [19, 77], including bridging the domain shift between real and generated images. Most relevant to our work, Isola et al. [24] propose an image-to-image translation framework using paired data, while CycleGAN [77] can do so with unpaired data. Our method also uses a image-to-image translation networks, but compared to CycleGAN our use of conditional image generation addresses the logical fallacy that an image-like label (a skeleton) does not contain sufficient information to generate a full image — this issue is discussed in depth in section 4.

Appearance and geometry factorization. Recent methods for image generation conditioned on object attributes, like viewpoint [44], pose [60], and hierarchical latents [52] have been proposed. Our method allows for similar but more fine-grained conditional image generation, conditioned on an appearance image or object landmarks. Many unsupervised methods for pose estimation [25, 33, 50, 67, 74] share similar ability. However, we can achieve more accurate and predictable image editing by manipulating semantic parts in the image through their corresponding landmarks.

3. Method

Our goal is to learn a network $\Phi : x \mapsto y$ that maps an image $x$ containing an object to its pose $y$. To avoid having to use image annotations, the network is trained using an auto-encoder formulation. Namely, given the pose $y = \Phi(x)$ extracted from the image, we train a decoder $\Psi$ that reconstructs the image from the pose. However, since pose lacks appearance information, this reconstruction task is ill posed. Hence, we also provide the decoder with a different image $x'$ of the same object to convey its appearance. Formally, the image $x$ is reconstructed from the pose $y$ and the auxiliary image $x'$ via a conditional decoder network

$$x = \Psi(\Phi(x), x').$$

Unfortunately, without additional constraints, this formulation fails to learn pose properly [25]. The reason is that, given enough freedom, the encoder $\Phi(x)$ may simply decide to output a copy of the input image $x$, which allows it to trivially satisfy constraint (1) without learning anything useful (this issue is visualized in section 4 and fig. 4). The formulation needs a mechanism to force the encoder $\Phi$ to ‘distil’ only pose information and discard appearance.

We make two key contributions to address these issues. First, we introduce a dual representation of pose as a vector of 2D keypoint coordinates and as a pictorial representation in the form of ‘skeleton’ image (section 3.1). We show that this dual representation provides a tight bottleneck that

$$\mathbf{\beta} \circ \eta \mathbf{y}$$

is used to drop appearance information that may leak in the pose image $\mathbf{y}$. A discriminator $\mathbf{D}$ is used to match the distribution of predicted poses to a reference prior distribution, represented by unpaired pose samples $\mathbf{\bar{y}}$. 

Figure 2. Architecture. We learn an encoder $\Phi$ that maps an image $x$ to its pose $y$, represented as a skeleton image. This is done via conditional auto-encoding, learning also a decoder $\Psi$ that reconstruct the input $x$ from its pose $y$ and a second auxiliary video frame $x'$. A bottleneck $\beta \circ \eta \mathbf{y}$ is used to drop appearance information that may leak in the pose image $\mathbf{y}$. A discriminator $\mathbf{D}$ is used to match the distribution of predicted poses to a reference prior distribution, represented by unpaired pose samples $\mathbf{\bar{y}}$.

Figure 3. Training data flow. Data flowing through our model (fig. 2) during training on the Human3.6M (human pose) and VoxCeleb2 (face) datasets. $\mathbf{y}$, $\mathbf{y}^*$ are our predictions.
The differentiable function $y$, the encoder/pose recogniser can be written as an image-to-image translation network $\Phi : x \mapsto y$ whose input $x \in \mathbb{R}^{3 \times H \times W}$ and output $y$ are both images. The same is true for the conditional decoder $\Psi : (y, x') \mapsto x$ of eq. $(1)$.

While image-to-image translation is desirable architecturally, the downside of encoding pose as an image $y$ is that it gives the encoder $\Phi$ an opportunity to ‘cheat’ and inject appearance information in the pose representation $y$. We can prevent cheating by exploiting the coordinate representation of pose to filter out any hidden appearance information form $y$. We do so by converting the pose image into keypoints and then back. This amounts to substituting $y = \beta \circ \eta(y)$ in eq. $(1)$, which yields the modified auto-encoding constraint:

$$x = \Psi(\beta \circ \eta \circ \Phi(x), x').$$  

### 3.2. Learning formulation & pose prior

**Auto-encoding loss.** In order to learn the auto-encoder $(3)$, we use a dataset of $N$ example pairs of video frames $\{(x_i, x'_i)\}_{i=1}^N$. Then the auto-encoding constraint $(3)$ is enforced by optimizing a reconstruction loss. Here we use a perceptual loss:

$$\mathcal{L}_{\text{perc}} = \frac{1}{N} \sum_{i=1}^N \left\| \Gamma(\hat{x}_i) - \Gamma(x_i) \right\|_2^2,$$

where $\hat{x}_i = \Psi(\beta \circ \eta \circ \Phi(x_i), x'_i)$ is the reconstructed image, $\Gamma$ is a feature extractor. Instead of comparing pixels directly, the perceptual loss compares features extracted from a standard network such as VGG [5, 11, 15, 26], and leads to more robust training.

**Pose prior.** In addition to the $N$ training image pairs $\{(x_i, x'_i)\}_{i=1}^N$, we also assume to have access to $M$ sample poses $\{p_j\}_{j=1}^M$. Importantly, these sample poses are unpaired, in the sense that they are not annotations of the training images.

We use the unpaired pose samples to encourage the predicted poses $y$ to be plausible. This is obtained by matching two distributions. The reference distribution $q(y)$ is given by the unpaired pose samples $\{y_j = \beta(p_j)\}_{j=1}^M$. The other distribution $p(y)$ is given by the pose samples $\{y_i = \Phi(x_i)\}_{i=1}^N$ predicted by the learned encoder network from the example video frames $x_i$.

The goal is to match $p(y) \approx q(y)$ in a distributional sense. This can be done by learning a discriminator network $D(y) \in [0,1]$ whose purpose is to discriminate between the unpaired samples $y_j = \beta(p_j)$ and the predicted samples $y_i = \Phi(x_i)$. Samples are compared by means of the difference adversarial loss of [34]:

$$\mathcal{L}_{\text{disc}}(D) = \frac{1}{M} \sum_{j=1}^M D(y_j)^2 + \frac{1}{N} \sum_{i=1}^N (1 - D(y_i))^2.$$

In addition to capturing plausible poses, the pose discriminator \( D(y) \) also encourages the images \( y \) to be ‘skeleton-like’. The effect is thus similar to the bottleneck introduced in section 3.1 and one may wonder if the discriminator makes the bottleneck redundant. The answer, as shown in sections 4 and 5, is negative: both are needed.

**Overall learning formulation.** Combining losses (4) and (5) yields the overall objective:

\[
L(\Phi, \Psi, D) = \lambda L_{\text{disc}}(D, \Phi) + L_{\text{perc}}(\Psi, \Phi),
\]

where \( \lambda \) is a loss-balancing factor. The components of this model and their relations are illustrated in fig. 2. Similar to any adversarial formulation, eq. (6) is minimized w.r.t. \( \Phi, \Psi \) and maximised w.r.t. \( D \).

**Details.** The functions \( \Phi, \Psi, \eta \) and \( D \) are implemented as convolutional neural networks. The auto-encoder functions \( \Phi \) and \( \Psi \) and the discriminator \( D \) are trained by optimizing the objective in eq. (6) (\( \eta \) is pre-trained using unpaired landmarks, for details see supplementary). Batches are formed by sampling random pairs of video frames \( (x_i, x'_i) \) and unpaired pose \( y_j \) samples. When sampling from image datasets (instead of videos), we generate image pairs as \((y_1(x_i), g_2(x_i))\) by applying random thin-plate-splines \( g_1, g_2 \) to training samples \( x_i \). All the networks are trained from scratch. Architectures and training details are in the supplementary.

4. Relation to image-to-image translation

Our method is related to unpaired image-to-image translation, of which CycleGAN [77] is perhaps the best example, but with two key differences: (a) it has a bottleneck (section 3.1) that prevents leaking appearance information into the pose representation \( y \), and (b) it reconstructs the image \( x \) conditioned on a second image \( x' \). We show in the experiments that these changes are critical for pose recognition performance, and conduct a further analysis here.

First, consider what happens if we drop both changes (a) and (b), thus making our formulation more similar to CycleGAN. In this case, eq. (1) reduces to \( x = \Psi(\Phi(x)) \). The trivial solution of setting both \( \Phi \) and \( \Psi \) to the identity functions is only avoided due to the discriminator loss (5), which encourages \( y = \Phi(x) \) to look like a skeleton (rather than a copy of \( x \)). In theory, then, this problem should be ill-posed as the pose \( y \) should not have sufficient information to recover the input image \( x \). However, the reconstructions from such a network still look reasonably good (see fig. 4). A closer look at logarithm of the generated skeleton \( y \), reveals that CycleGAN ‘cheats’ by leaking appearance information via subtle patterns in \( y \). By contrast, our bottleneck significantly limits leaking appearance in the pose image and thus its ability to reconstruct \( x = \Psi(\beta \circ \eta \circ \Phi(x)) \) from a single image; instead, reconstruction is achieved by injecting the missing appearance information via the auxiliary image \( x' \) using a conditional image decoder (eq. (3)).

5. Experiments

We evaluate our method on the task of 2D landmark detection for human pose (section 5.1), faces (section 5.2), and cat heads (section 5.3) and outperform state-of-the-art methods (tables 1 to 3) on these tasks. We examine the relative contributions of components of our model in an ablation study (section 5.4). We study the effect of reducing the number of pose samples used in the empirical prior (section 5.5). Finally, we demonstrate image generation and manipulation conditioned on appearance and pose (section 5.6).

**Evaluation.** Our method directly outputs predictions for keypoints that are human-interpretible. In contrast, self-supervised methods [25, 33, 54, 55, 56, 67, 74] predict only machine-interpretible keypoints, as illustrated in fig. 1, and require at least some example images with paired keypoint annotations in order to learn to convert these landmarks to human-interpretible ones for benchmarking or for applications. We call this step supervised post-processing. Our method does not require this step, but we also include this result for a direct comparison with previous methods.

5.1. Human pose

**Datasets.** *Simplified Human3.6M* introduced by Zhang et al. [74] for evaluating unsupervised pose recognition, contains 6 activities in which human bodies are mostly upright; it comprises 800k training and 90k testing images. *Human3.6M* [23] is a large-scale dataset that contains 3.6M accurate 2D and 3D human pose annotations for 17 different activities, imaged under 4 viewpoints and a static background. For training, we use subjects 1, 5, 6, 7, and 8, and subjects 9 and 11 for evaluation, as in [64]. *PennAction* [73]
contains 2k challenging consumer videos of 15 sports categories. *MPI-INF-3DHP* [35] is a mocap dataset containing 8 subjects performing 8 activities in complex exercise poses. There are 28 joints annotated.

We split datasets into two disjoint parts for sampling image pairs \((x, x')\) (cropped to the provided bounding boxes), and skeleton prior respectively to ensure that the pose data does not contain labels corresponding to the training images. For the Human3.6M datasets we split the videos in half, while for PennAction we split in half the set of videos from each action category. We also evaluate the case when images and skeletons are sampled from different datasets and for this purpose we use the MPI-INF-3DHP mocap data.

**Evaluation.** We report 2D landmark detection performance on the simplified and original Human3.6M datasets. For Simplified Human3.6M, we follow the standard protocol of [74] and report the error for all 32 joints normalized by the image size. For Human3.6M, we instead report the mean error in pixels over 17 of the 32 joints [23]. To demonstrate learning from unpaired prior, we consider two settings for sourcing the images and the prior. In the first setting, we use different datasets for the two, and sample images from Human3.6M and poses from MPI-INF-3DHP. In the second setting, we use instead two disjoint parts of the same dataset Human3.6M for both images and poses. When using MPI-INF-3DHP dataset as the prior, we predict 28 joints, but use 17 joints that are common with Human3.6M for evaluation. We train our method from scratch and compare its performance with both supervised and unsupervised methods.

**Results.** Table 1 reports the results on Simplified Human3.6M. As in previous self-supervised works [54, 74], we compare against the supervised baseline by Newell et al. [36]. Our model outperforms all the baselines [33, 54, 74] without the supervised post-processing used by the others.

Table 1. Human landmark detection (Simplified H3.6M). Comparison with state-of-the-art methods for human landmark detection on the Simplified Human3.6M dataset [74]. We report % -MSE normalised by image size for each activity.

<table>
<thead>
<tr>
<th>Method</th>
<th>all</th>
<th>wait pose greet direct discuss walk</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>fully supervised</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Newell et al. [36]</td>
<td>2.16</td>
<td>1.88</td>
</tr>
<tr>
<td><strong>self-supervised + supervised post-processing</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thewlis et al. [54]</td>
<td>7.51</td>
<td>7.54</td>
</tr>
<tr>
<td>Zhang et al. [74]</td>
<td>4.14</td>
<td>5.01</td>
</tr>
<tr>
<td>Lorenz et al. [33]</td>
<td>2.79</td>
<td>—</td>
</tr>
<tr>
<td><strong>self-supervised (no post-processing)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>2.73</td>
<td>2.66</td>
</tr>
</tbody>
</table>

Table 2. Human landmark detection (full H3.6M). Comparison on Human3.6M test set with a supervised baseline Newell et al. [36], and a self-supervised method [25]. We report the MSE in pixels [23]. Results for each activity are in the supplementary.

<table>
<thead>
<tr>
<th>Method</th>
<th>Human3.6M</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>fully supervised</strong></td>
<td></td>
</tr>
<tr>
<td>Newell et al. [36]</td>
<td>19.52</td>
</tr>
<tr>
<td><strong>self-supervised + supervised post-processing</strong></td>
<td></td>
</tr>
<tr>
<td>Jakab &amp; Gupta et al. [25]</td>
<td>19.12</td>
</tr>
<tr>
<td><strong>self-supervised (no post-processing)</strong></td>
<td></td>
</tr>
<tr>
<td>Ours with 3DHP prior</td>
<td>18.94</td>
</tr>
<tr>
<td>Ours with H3.6M prior</td>
<td>14.46</td>
</tr>
</tbody>
</table>

It may be surprising that our method outperforms the supervised baseline. A possible reason is the limited number of supervised examples, which causes the supervised baseline to overfit. This can be noted by comparing the training / test errors: 14.61 / 19.52 for supervised hourglass and 13.79 / 14.46 for our method.

When poses are sampled from a different dataset (MPI-INF-3DHP) than the images (Human3.6M), the error is higher at 18.94 (but still better than the supervised alternative). This increase is due to the domain gap between the two datasets. Figure 5 shows some qualitative examples. Limitations of the method are highlighted in fig. 6.

![Figure 6. Limitations.](image)

**Figure 6. Limitations.** [1-2] complex human poses like sitting are challenging to learn from a weak pose prior, [3] it could be difficult to disambiguate the sides due to bilateral symmetry, [4-5] occlusions are difficult to handle.

### 5.2. Human faces

**Datasets.** *VoxCeleb2* [10] is a large-scale dataset consisting of 1M short clips of talking-head videos extracted from YouTube. *MultiPIE* [51] contains 68 labelled facial landmarks and 6k samples. We use this dataset as the only source for the prior. *300-W* [47] is a challenging dataset of facial images obtained by combining multiple datasets [4, 42, 76] as described in [43, 54]. As in MultiPIE, 300-W contains 68 annotated facial landmarks. We use 300-W as our test dataset and follow the evaluation protocol in [43].

**Results.** As for human pose, we study a scenario where images and poses are sourced from a different dataset. Using VoxCeleb2 and 300-W for the images, and MultiPIE (6k samples) for the poses (fig. 7). We train our method from scratch using video frames from VoxCeleb2; then we
fine-tune the model using our unsupervised method on the 300-W training images. We report performance on 300-W test set in table 3. Our method performs well even without any supervised fine-tuning on the target 300-W, and it already outperforms the unsupervised method of [55]. Adding supervised post-processing (on 300-W training set) as done in all self-supervised learning methods [54, 55, 56, 67], we outperform all except for [56] when they use their HG network that has 3 times more learnable parameters (4M vs 12M parameters). Interestingly we also outperform all supervised methods except [13, 68].

![Unpaired transfer](image)

Figure 7. Unpaired transfer. We leverage approx. 6k landmarks from the MultiPIE dataset [51] as a prior [top] and unlabelled images from the large-scale VoxCeleb2 [10] [middle] (1M clips, 6k identities) to train a detector that we test on the 300-W dataset [47] [bottom] (predictions in green) with state-of-the-art results (table 3). More qualitative results are in the supplementary.

5.3. Cat heads

Cat Head [72] dataset contains 9k images of cat heads each annotated with 7 landmarks. We use the same train and test split as [74]. We split the training set into two equally sized parts with no overlap. The first one is used to sample training images and the second one for the landmark prior. Our predictions are visualized in fig. 8.

![Cat head landmarks](image)

Figure 8. Cat head landmarks. Our predictions on Cat Head test set [72] consistently track landmarks across different views. More results are included in the supplementary.

5.4. Ablation study

As noted above, we can obtain our method by making the following changes to CycleGAN: (1) switching to a conditional image generator \( \Psi \), (2) introducing the skeleton bottleneck \( \beta \circ \eta \), and (3) removing the “second auto-encoder cycle” for the other domain (in our case the skeleton images). Table 4 shows the effect of modifying CycleGAN in this manner on Simplified Human3.6M [74] for humans and on 300-W [47] for faces.

The baseline CycleGAN can be thought of as learning a mapping between images and skeletons via off-the-shelf image translation. Switching to a conditional image generator (1) does not improve the results because the model can still leak appearance information’s pose. However, introducing the bottleneck (2) improves performance significantly for both humans (2.86% vs. 3.54% CycleGAN, a 20% error reduction) and faces (11.89% vs. 9.64% CycleGAN, a 19% error reduction). This also justifies the use of a conditional generator as the model fails to converge if the bottleneck is used without it. Removing the second cycle (3) leads to further improvements, showing that this part is detrimental for our task.

5.5. Unpaired sample efficiency

Table 5 demonstrates that our method retains state-of-the-art performance even when we use only 50 unpaired landmark samples for the empirical prior. The experiment was done following the same protocols for training on face and human datasets as described previously.

<table>
<thead>
<tr>
<th>Method</th>
<th>300-W</th>
</tr>
</thead>
<tbody>
<tr>
<td>fully supervised</td>
<td></td>
</tr>
<tr>
<td>LBF [43]</td>
<td>6.32</td>
</tr>
<tr>
<td>CFSS [78]</td>
<td>5.76</td>
</tr>
<tr>
<td>cGPRT [30]</td>
<td>5.71</td>
</tr>
<tr>
<td>DDN [71]</td>
<td>5.65</td>
</tr>
<tr>
<td>TCDCN [75]</td>
<td>5.54</td>
</tr>
<tr>
<td>RAR [68]</td>
<td>4.94</td>
</tr>
<tr>
<td>self-supervised + supervised post-processing</td>
<td></td>
</tr>
<tr>
<td>Thewlis et al. [55]</td>
<td>9.30</td>
</tr>
<tr>
<td>Thewlis et al. [54]</td>
<td>7.97</td>
</tr>
<tr>
<td>Thewlis et al. [56] SmallNet ( \dagger )</td>
<td>5.75</td>
</tr>
<tr>
<td>Wiles et al. [67]</td>
<td>5.71</td>
</tr>
<tr>
<td>Jakab &amp; Gupta et al. [25]</td>
<td>5.39</td>
</tr>
<tr>
<td>Thewlis et al. [56] HourGlass ( \dagger )</td>
<td>4.65</td>
</tr>
<tr>
<td>self-supervised</td>
<td></td>
</tr>
<tr>
<td>Ours (no post-processing)</td>
<td>8.67</td>
</tr>
<tr>
<td>+ supervised post-processing</td>
<td>5.12</td>
</tr>
</tbody>
</table>

Table 3. Facial landmark detection. Comparison with state-of-the-art methods on 2D facial landmark detection. We report the inter-ocular distance normalised keypoint localisation error [75] (in %; ↓ is better) on the 300-W test set. \( \dagger \): [56] evaluate using two different networks: (1) SmallNet which we outperform, (2) HourGlass is not directly comparable due to much larger capacity (4M vs 12M parameters).
Though generating high quality images is not our primary task (last row). We report 2D landmark detection error (\(\cdot\)) of eq. (1) can also be used to produce novel images by combining pose and appearance from different images. Figure 9 shows that the model can be used to transfer the appearance of a human face identity on top of the pose of another. Though generating high quality images is not our primary goal, the ability to transfer appearance shows that our method properly factorizes the latter from pose.

This also demonstrates significant generalization over the training setting, as the system only learns from pairs of frames sampled from the same video and thus with same identity, but it can swap different identities. In fig. 10, we further leverage the disentanglement of geometry and appearance to manipulate a face by editing its keypoints.

### 5.6. Appearance and geometry factorization

The conditional image generator \(\Psi : (y'; x') \rightarrow \hat{x}\) of eq. (1) can also be used to produce novel images by combining pose and appearance from different images. Figure 9 shows that the model can be used to transfer the appearance of a human face identity on top of the pose of another. Though generating high quality images is not our primary goal, the ability to transfer appearance shows that our method properly factorizes the latter from pose.

This also demonstrates significant generalization over the training setting, as the system only learns from pairs of frames sampled from the same video and thus with same identity, but it can swap different identities. In fig. 10, we further leverage the disentanglement of geometry and appearance to manipulate a face by editing its keypoints.

### 6. Conclusion

We have shown that combining conditional image generation with a dual representation of pose with a tight geometric bottleneck can be used to learn to recognize the pose of complex objects such as humans without providing any labelled image to the system. In order to do so, our method makes use of an unpaired pose prior, which also allows it to output human-interpretable pose parameters. With this, we have achieved optimal landmark detection accuracy for methods that do not use labelled images for training.

**Acknowledgements.** We are grateful for the support of ERC 638009-IDIU, and the Clarendon Fund Scholarship. We would like to thank Triantafyllos Afouras, Relja Arandjelović, and Chuhan Zhang for helpful advice.

**References**


[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and...


