Self-supervised Learning of Interpretable Keypoints from Unlabelled Videos

OPEN ACCESS

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 distils 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/
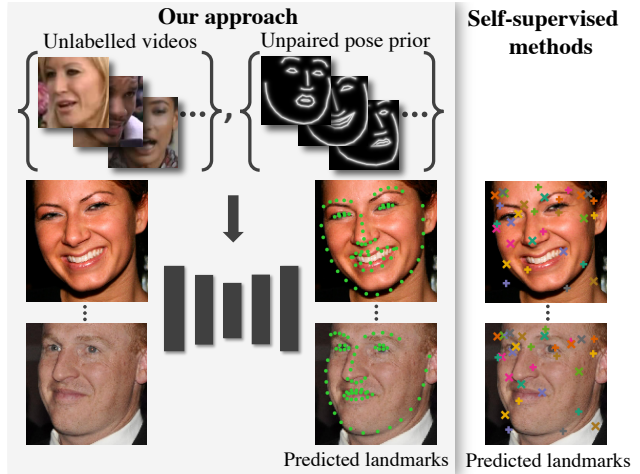
Figure 1. **Learning landmark detectors from unpaired data.** We learn to directly predict human-interpretable landmarks of an object using only unlabelled videos and a prior on the possible landmark configurations **[left]**. The prior can be obtained from unpaired supervision or from a different modality, such as mocap data. Our method obtains state-of-the-art landmark detection performance for approaches that use unlabelled images for supervision. In contrast, self-supervised landmark detectors [25, 33, 54, 74] can only learn to discover keypoints **[right]** that are not human-interpretable (predictions from [25]) and require supervised post-processing.

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 for-

mulate 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. By contrast, state-of-the-art self-supervised keypoint detectors [25, 50, 54, 67, 74] do not learn "semantic" keypoints and, in post-processing, they need at least some *paired supervision* to output human-interpretable keypoints. We highlight this difference in fig. 1.

Overall, we make three significant contributions:

1. We introduce a new conditional generator design combining image translation, a new bottleneck using a dual representation of pose, and an adversarial loss which significantly improve recognition performance.
2. We learn, for the first time, to directly predict human-interpretable landmarks without requiring any labelled images.
3. We obtain state-of-the-art unsupervised landmark detection performance even when compared against methods that use paired supervision in post-processing.

We test our approach using videos of people, faces, and cat images. On standard benchmarks such as Human3.6M [23] and 300-W [47], we achieve state-of-the-art pose recognition performance for methods that learn only from unlabelled images. We also probe generalization by testing whether the empirical pose prior can be extracted independently from the videos used to train the pose recognizer. We demonstrate this in two challenging scenarios. First, we use the mocap data from MPI-INF-3DHP [35] as prior and we learn a human pose recognizer on videos from Human3.6M. Second, we use the MultiPIE [51] dataset as prior to learn a face pose recognizer on VoxCeleb2 [10] videos, and achieve state-of-the-art facial keypoint detection performance on 300-W.

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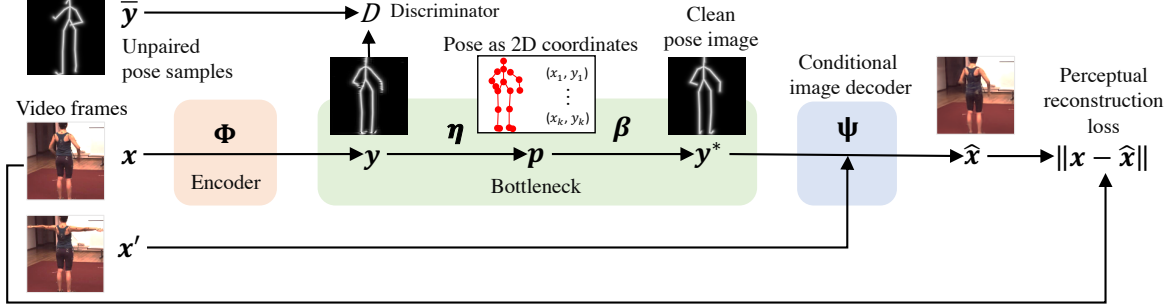
Figure 2. **Architecture.** We learn an encoder Φ that maps an image $\boldsymbol{x}$ to its pose $\boldsymbol{y}$, represented as a skeleton image. This is done via conditional auto-encoding, learning also a decoder Ψ that reconstruct the input $\boldsymbol{x}$ from its pose $\boldsymbol{y}$ and a second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

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 $\Phi : \boldsymbol{x} \mapsto \boldsymbol{y}$ that maps an image $\boldsymbol{x}$ containing an object to its pose $\boldsymbol{y}$. To avoid having to use image annotations, the network is trained using an
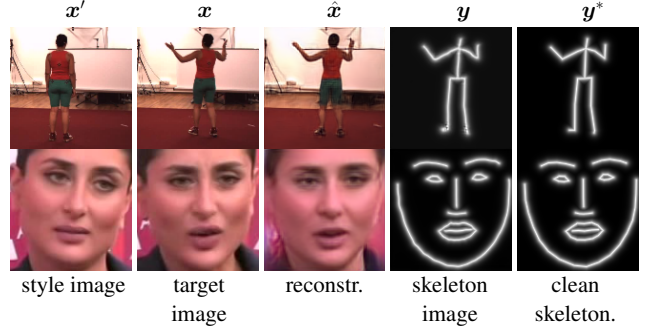

Figure 3. **Training data flow.** Data flowing through our model (fig. 2) during training on the Human3.6M (human pose) and VoxCeleb2 (face) datasets. $\boldsymbol{y}, \boldsymbol{y}^*$ are our predictions.

auto-encoder formulation. Namely, given the pose $\boldsymbol{y} = \Phi(\boldsymbol{x})$ extracted from the image, we train a decoder network Ψ 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 $\boldsymbol{x}'$ of the same object to convey its appearance. Formally, the image $\boldsymbol{x}$ is reconstructed from the pose $\boldsymbol{y}$ and the auxiliary image $\boldsymbol{x}'$ via a *conditional decoder network*

$$\boldsymbol{x} = \Psi(\Phi(\boldsymbol{x}), \boldsymbol{x}'). \tag{1}$$

Unfortunately, without additional constraints, this formulation fails to learn pose properly [25]. The reason is that, given enough freedom, the encoder $\Phi(\boldsymbol{x})$ may simply decide to output a copy of the input image $\boldsymbol{x}$, 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 Φ 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
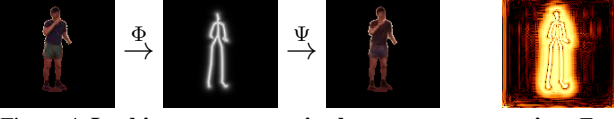
Figure 4. **Leaking appearance in the pose representation.** From left to right: input image $\boldsymbol{x}$, extracted skeleton image $\boldsymbol{y} = \Phi(\boldsymbol{x})$, and image reconstruction $\hat{\boldsymbol{x}} = \Psi(\Phi(\boldsymbol{x}))$. In principle, it should not be possible to reconstruct the full image from only the skeleton, but the function Φ can 'hide' the necessary information in a structured noise pattern, shown to the right as $\log \Phi(\boldsymbol{x})$.

distils pose information effectively while making it possible to implement the auto-encoder (1) using powerful image-to-image translation networks.

Our second contribution is to introduce an *empirical prior* on the possible object poses (section 3.2). In this manner, we can constrain not just the individual pose samples $\boldsymbol{y}$, but their *distribution* $p(\boldsymbol{y})$ as well. In practice, the prior allows to use unpaired pose samples to improve accuracy and to learn an human-interpretable notion of pose that does not necessitate further learning to be used in applications.

3.1. Dual representation of pose & bottleneck

We consider a dual representation of the pose of an object as a vector of K 2D *keypoint coordinates* $\boldsymbol{p} = (p_1, \ldots, p_K) \in \Omega^K$ and as an *image* $\boldsymbol{y} \in \mathbb{R}^\Omega$ containing a pictorial rendition of the pose as a skeleton (see fig. 2 for an illustration). Here the symbol $\Omega = \{1, \ldots, H\} \times \{1, \ldots, W\}$ denotes a grid of pixel coordinates.

Representing pose as a set of 2D keypoints provides a tight bottleneck that preserves geometry but discards appearance information. Representing pose as a skeleton image allows to implement the encoder and decoder networks as image translation networks. In particular, the image of the object $\boldsymbol{x}$ and of its skeleton $\boldsymbol{y}$ are *spatially aligned*, which makes it easier for a CNN to map between them.

Next, we show how to switch between the two representations of pose. We define the mapping $\boldsymbol{y} = \beta(\boldsymbol{p})$ from the coordinates $\boldsymbol{p}$ to the skeleton image $\boldsymbol{y}$ *analytically*. Let E be the set of keypoint pairs (i, j) connected by a skeleton edge and let $u \in \Omega$ be an image pixel. Then the skeleton image is given by:

$$\beta(\boldsymbol{p})_u = \exp\left(-\gamma \min_{(i,j) \in E, r \in [0,1]} \|u - r\boldsymbol{p}_i - (1-r)\boldsymbol{p}_j\|^2\right) \tag{2}$$

The differentiable function $\boldsymbol{y} = \beta(\boldsymbol{p})$ defines a distance field from line segments that form the skeleton and applies an exponential fall off to generate an image. The visual effect is to produce a smooth line drawing of the skeleton. We also train an inverse function $\boldsymbol{p} = \eta(\boldsymbol{y})$, implementing it as a neural network regressor (see supplementary for details).

Given the two maps (η, β), we can use either representation of pose, as needed. In particular, by using the pictorial

representation $\boldsymbol{y}$, the encoder/pose recogniser can be written as an image-to-image translation network $\Phi : \boldsymbol{x} \mapsto \boldsymbol{y}$ whose input $\boldsymbol{x} \in \mathbb{R}^{3 \times H \times W}$ and output $\boldsymbol{y}$ are both images. The same is true for the conditional decoder $\Psi : (\boldsymbol{y}, \boldsymbol{x}') \mapsto \boldsymbol{x}$ of eq. (1).

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

$$\boldsymbol{x} = \Psi(\beta \circ \eta \circ \Phi(\boldsymbol{x}), \boldsymbol{x}'). \tag{3}$$

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 $\{(\boldsymbol{x}_i, \boldsymbol{x}'_i)\}_{i=1}^N$. Then the auto-encoding constraint (3) is enforced by optimizing a reconstruction loss. Here we use a *perceptual loss*:

$$\mathcal{L}_{\text{perc}} = \frac{1}{N} \sum_{i=1}^N \|\Gamma(\hat{\boldsymbol{x}}_i) - \Gamma(\boldsymbol{x}_i)\|_2^2, \tag{4}$$

where $\hat{\boldsymbol{x}}_i = \Psi(\beta \circ \eta \circ \Phi(\boldsymbol{x}_i), \boldsymbol{x}'_i)$ is the reconstructed image, Γ 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 $\{(\boldsymbol{x}_i, \boldsymbol{x}'_i)\}_{i=1}^N$, we also assume to have access to M sample poses $\{\bar{\boldsymbol{p}}_j\}_{j=1}^M$. 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 $\boldsymbol{y}$ to be plausible. This is obtained by matching two distributions. The reference distribution $q(\boldsymbol{y})$ is given by the unpaired pose samples $\{\bar{\boldsymbol{y}}_j = \beta(\bar{\boldsymbol{p}}_j)\}_{j=1}^M$. The other distribution $p(\boldsymbol{y})$ is given by the pose samples $\{\boldsymbol{y}_i = \Phi(\boldsymbol{x}_i)\}_{i=1}^N$ predicted by the learned encoder network from the example video frames $\boldsymbol{x}_i$.

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

$$\mathcal{L}_{\text{disc}}(D) = \frac{1}{M} \sum_{j=1}^M D(\bar{\boldsymbol{y}}_j)^2 + \frac{1}{N} \sum_{i=1}^N (1 - D(\boldsymbol{y}_i))^2. \tag{5}$$

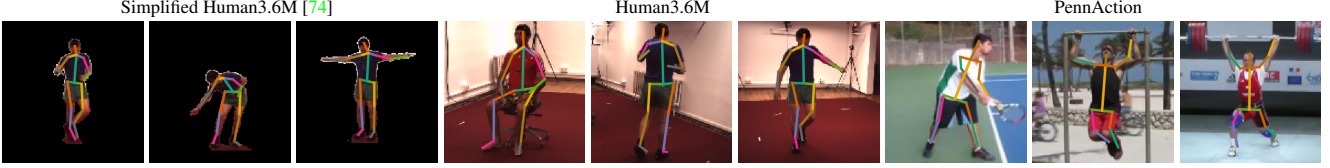| Simplified Human3.6M [74] | Human3.6M | PennAction |

Figure 5. **Human pose predictions.** 2D keypoint predictions (visualised as connected limbs) on the simplified [74] (with no background), full Human3.6M [23], and PennAction [73] test sets. Our method directly predicts human landmarks in complex poses without any additional supervision. More samples are included in the supplementary.

In addition to capturing plausible poses, the pose discriminator $D(\boldsymbol{y})$ *also* encourages the images $\boldsymbol{y}$ 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:

$$\mathcal{L}(\Phi, \Psi, D) = \lambda \mathcal{L}_{\text{disc}}(D, \Phi) + \mathcal{L}_{\text{perc}}(\Psi, \Phi), \qquad (6)$$

where λ 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. Φ, Ψ and maximised w.r.t. D.

Details. The functions Φ, Ψ, η and D are implemented as convolutional neural networks. The auto-encoder functions Φ and Ψ and the discriminator D are trained by optimizing the objective in eq. (6) (η is pre-trained using unpaired landmarks, for details see supplementary). Batches are formed by sampling random pairs of *video frames* $(\boldsymbol{x}_i, \boldsymbol{x}_i')$ and unpaired pose $\bar{\boldsymbol{y}}_j$ samples. When sampling from *image datasets* (instead of videos), we generate image pairs as $(g_1(\boldsymbol{x}_i), g_2(\boldsymbol{x}_i))$ by applying random thin-plate-splines g_1, g_2 to training samples $\boldsymbol{x}_i$. 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 $\boldsymbol{y}$, and (b) it reconstructs the image $\boldsymbol{x}$ conditioned on a second image $\boldsymbol{x}'$. 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 $\boldsymbol{x} = \Psi(\Phi(\boldsymbol{x}))$. The trivial solution of setting both Φ and Ψ to the identity functions is only avoided due to the discriminator loss (5), which encourages $\boldsymbol{y} = \Phi(\boldsymbol{x})$ to look like a skeleton (rather than a copy of $\boldsymbol{x}$). In theory, then, this problem should be

ill-posed as the pose $\boldsymbol{y}$ should not have sufficient information to recover the input image $\boldsymbol{x}$. However, the reconstructions from such a network still look reasonably good (see fig. 4). A closer look at logarithm of the generated skeleton $\boldsymbol{y}$, reveals that CycleGAN 'cheats' by leaking appearance information via subtle patterns in $\boldsymbol{y}$. By contrast, our bottleneck significantly limits leaking appearance in the pose image and thus its ability to reconstruct $\boldsymbol{x} = \Psi(\beta \circ \eta \circ \Phi(\boldsymbol{x}))$ from a single image; instead, reconstruction is achieved by injecting the missing appearance information via the auxiliary image $\boldsymbol{x}'$ 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-interpretable. In contrast, self-supervised methods [25, 33, 54, 55, 56, 67, 74] predict only *machine-interpretable* 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-interpretable 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]

Method	all	wait	pose	greet	direct	discuss	walk
fully supervised							
Newell *et al.* [36]	**2.16**	**1.88**	**1.92**	**2.15**	**1.62**	**1.88**	**2.21**
self-supervised + supervised post-processing							
Thewlis *et al.* [54]	7.51	7.54	8.56	7.26	6.47	7.93	5.40
Zhang *et al.* [74]	4.14	5.01	4.61	4.76	4.45	4.91	4.61
Lorenz *et al.* [33]	2.79	—	—	—	—	—	—
self-supervised (no post-processing)							
Ours		2.73	2.66	2.27	2.73	2.35	2.35	4.00

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.

Method	Human3.6M
fully supervised	
Newell *et al.* [36]	19.52
self-supervised + supervised post-processing	
Jakab & Gupta *et al.* [25]	19.12
self-supervised (no post-processing)	
Ours with *3DHP prior*	18.94
Ours with *H3.6M prior*	**14.46**

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.

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 2 summarises our results on the original Human3.6M test set. Here we also compare against the supervised baseline [36] and the self-supervised method of [25]. Our model outperforms the baselines in this test too.

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.
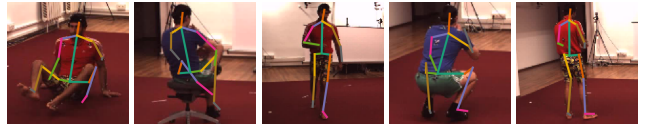

Figure 6. **Limitations.** **[1-2]** complex human poses like sitting are challenging to learn from a weak pose prior, **[3]** it could be difficult to disambiguate the sides due to bilateral symmetry, **[4-5]** occlusions are difficult to handle.

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* datasets, 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].
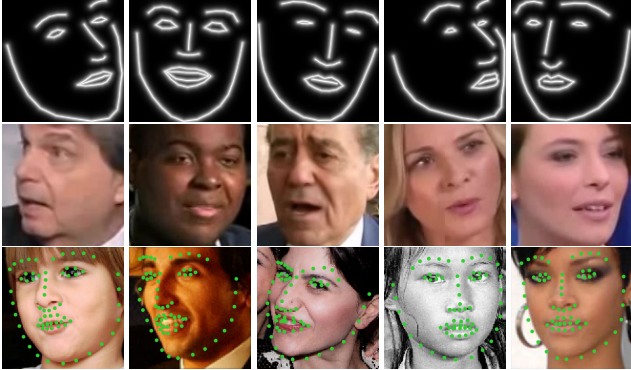

Figure 7. **Unpaired transfer.** We leverage approx. 6k landmarks from the MultiPIE dataset [51] as a prior **[top]** and unlabelled images from the the large-scale VoxCeleb2 [10] **[middle]** (1M clips, 6k identities) to train a detector that we test on the 300-W dataset [47] **[bottom]** (predictions 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.

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 Ψ, (2) introducing the skeleton bottleneck $\beta \circ \eta$, and (3) removing the "second auto-encoder

Method	300-W
fully supervised	
LBF [43]	6.32
CFSS [78]	5.76
cGPRT [30]	5.71
DDN [71]	5.65
TCDCN [75]	5.54
RAR [68]	4.94
Wing Loss [13]	**4.04**
self-supervised + supervised post-processing	
Thewlis *et al*. [55]	9.30
Thewlis *et al*. [54]	7.97
Thewlis *et al*. [56] *SmallNet* [†]	5.75
Wiles *et al*. [67]	5.71
Jakab & Gupta *et al*. [25]	5.39
Thewlis *et al*. [56] *HourGlass* [†]	**4.65**
self-supervised	
Ours (no post-processing)	8.67
+ supervised post-processing	**5.12**

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. †: [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).

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.

Method	humans	faces
CycleGAN	3.54	11.89
+ conditional generator (1)	3.60	–
+ skeleton-bottleneck (2)	2.86	9.64
$-$ 2nd cycle = **ours** (3)	2.73	8.67
CycleGAN $-$ 2nd cycle	3.39	11.36

Table 4. **Ablation study.** We start with the CycleGAN [77] model and sequentially augment it with — (1) conditional image generator (Ψ), (2) skeleton bottleneck ($\beta \circ \eta$), and (3) remove the second cycle-constraint resulting in our proposed model. An auto-encoding model with a skeleton image as the intermediate representation (*i.e.* no keypoint bottleneck) and an adversarial loss is also reported (last row). We report 2D landmark detection error ($\downarrow$ is better) on the Simplified Human3.6M (section 5.1) for human pose, on the 300-W (section 5.2) for faces.

# unpaired samples	humans	faces	
	no post-proc.	*no post-proc.*	*+ sup. post-proc.*
full dataset	2.73	8.67	5.12
5000	2.92 ± 0.05	–	–
500	3.30 ± 0.06	8.91 ± 0.15	5.22 ± 0.04
50	4.05 ± 0.02	8.92 ± 0.20	5.19 ± 0.06

Table 5. **Varying # of unpaired landmark samples.** We train our method using varying numbers of samples for landmark prior. For faces, we sample the prior from MultiPIE dataset and evaluate on 300-W (section 5.2). For human pose, we sample the prior from the disjoint part of the Simplified Human3.6M training set and evaluate on the test set (section 5.1). We report the keypoint localisation error ($\pm\sigma$) (in %; $\downarrow$ is better). Full dataset has 6k unpaired samples for faces, and 400k for humans. Decreasing the number of unpaired landmark samples retains most of the performance.

5.6. Appearance and geometry factorization

The conditional image generator $\Psi : (\boldsymbol{y}^*, \boldsymbol{x}') \mapsto \hat{\boldsymbol{x}}$ 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

Figure 9. **Factorization of appearance and geometry.** *Reconstructed* image inherits appearance from the *style* image and geometry from the *target* image. **[left]:** human pose samples from Human3.6M. **[right]:** face samples from VoxCeleb2.
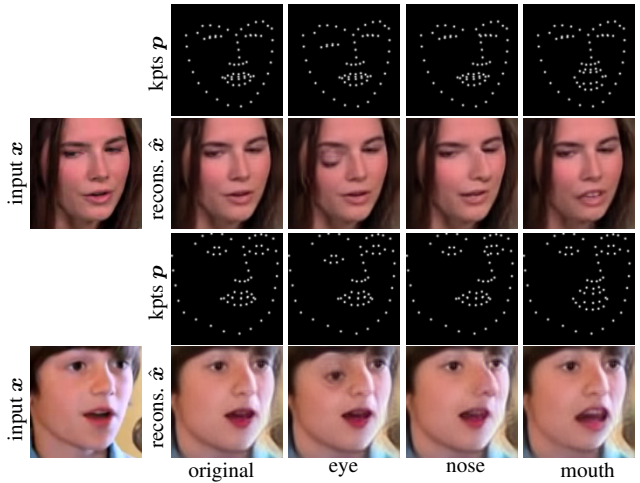

Figure 10. **Image editing using detected landmarks.** We show fine-grained control over the generated image by manipulating the coordinates of detected keypoints (*kpts*). The resulting changes are localised. Apart from demonstrating successful disentanglement of appearance and geometry, this also suggests that the model assigns correct semantics to the detected landmarks.

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.

References

[1] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Proc. CVPR*, pages 1014–1021. IEEE, 2009. 2

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

Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proc. CVPR*, pages 3686–3693, 2014. 2

[3] Vasileios Belagiannis and Andrew Zisserman. Recurrent human pose estimation. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 468–475. IEEE, 2017. 2

[4] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *TPAMI*, 35(12):2930–2940, 2013. 6

[5] Joan Bruna, Pablo Sprechmann, and Yann LeCun. Super-resolution with deep convolutional sufficient statistics. In *Proc. ICLR*, 2016. 4

[6] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *Proc. ECCV*, pages 717–732. Springer, 2016. 2

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proc. CVPR*, pages 7291–7299, 2017. 2

[8] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Proc. CVPR*, pages 4733–4742, 2016. 2

[9] Xianjie Chen and Alan L Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Proc. NIPS*, pages 1736–1744, 2014. 2

[10] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[11] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems*, pages 658–666, 2016. 4

[12] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International journal of computer vision*, 61(1):55–79, 2005. 2

[13] 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 *Proc. CVPR*, 2018. 7

[14] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015. 3

[15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proc. CVPR*, pages 2414–2423, 2016. 4

[16] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. GANFIT: Generative adversarial network fitting for high fidelity 3D face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[17] Zhenglin Geng, Chen Cao, and Sergey Tulyakov. 3D guided fine-grained face manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[18] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models - an open framework. In *Proc. Automatic Face & Gesture Recognition*, 2018.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. NIPS*, pages 2672–2680, 2014. 3

[20] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Learning to read by spelling: Towards unsupervised text recognition. In *Proc. ICVGIP*, 2018. 3

[21] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017. 3

[22] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *Proc. ECCV*, pages 34–50. Springer, 2016. 2

[23] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, jul 2014. 2, 5, 6

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proc. CVPR*, 2017. 3

[25] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Proc. NIPS*, 2018. 1, 2, 3, 5, 6, 7

[26] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. ECCV*, pages 694–711. Springer, 2016. 4

[27] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proc. CVPR*, pages 1465–1472. IEEE, 2011. 2

[28] Angjoo Kanazawa, David W Jacobs, and Manmohan Chandraker. Warpnet: Weakly supervised matching for single-view reconstruction. In *Proc. CVPR*, pages 3253–3261, 2016. 2

[29] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proc. CVPR*, 2018. 2

[30] Donghoon Lee, Hyunsin Park, and Chang D Yoo. Face alignment using cascade gaussian process regression trees. In *Proc. CVPR*, 2015. 7

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. ECCV*, pages 740–755. Springer, 2014. 2

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):248, 2015. 2

[33] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10955–10964, 2019. 1, 2, 3, 5, 6

[34] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proc. ICCV*, pages 2794–2802, 2017.

4

[35] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 International Conference on 3D Vision (3DV)*, pages 506–516. IEEE, 2017. 2, 6

[36] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proc. ECCV*, 2016. 2, 6

[37] Wanli Ouyang, Xiao Chu, and Xiaogang Wang. Multi-source deep learning for human pose estimation. In *Proc. CVPR*, pages 2329–2336, 2014. 2

[38] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *The IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2009. 2

[39] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *Proc. CVPR*, pages 1913–1921, 2015. 2

[40] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Poselet conditioned pictorial structures. In *Proc. CVPR*, pages 588–595, 2013. 2

[41] Varun Ramakrishna, Daniel Munoz, Martial Hebert, James Andrew Bagnell, and Yaser Sheikh. Pose machines: Articulated pose estimation via inference machines. In *Proc. ECCV*, pages 33–47. Springer, 2014. 2

[42] Deva Ramanan and Xiangxin Zhu. Face detection, pose estimation, and landmark localization in the wild. In *Proc. CVPR*, 2012. 6

[43] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *Proc. CVPR*, 2014. 6, 7

[44] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–767, 2018.

[45] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proc. CVPR*, volume 2, 2017. 2

[46] Matteo Ruggero Ronchi, Oisin Mac Aodha, Robert Eng, and Pietro Perona. It's all relative: Monocular 3d human pose estimation from weakly supervised data. In *BMVC*, 2018. 2

[47] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016. 2, 6, 7

[48] Benjamin Sapp, Chris Jordan, and Ben Taskar. Adaptive pose priors for pictorial structures. In *Proc. CVPR*, pages 422–429, 2010. 2

[49] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David Jacobs. Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[50] Zhixin Shu, Mihir Sahasrabudhe, Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proceedings of the European Conference on Computer Vision*, 2018. 2, 3

[51] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 53–58. IEEE, 2002. 2, 6, 7

[52] Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6490–6499, 2019.

[53] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *Proc. ECCV*, pages 712–729. Springer, 2018. 2

[54] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *Proc. ICCV*, 2017. 1, 2, 5, 6, 7

[55] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *Proc. NIPS*, 2017. 2, 5, 7

[56] James Thewlis, Samuel Albanie, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6361–6371, 2019. 2, 5, 7

[57] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proc. CVPR*, pages 648–656, 2015. 2

[58] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*, pages 1799–1807, 2014. 2

[59] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proc. CVPR*, pages 1653–1660, 2014. 2

[60] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1415–1424, 2017.

[61] Hsiao-Yu Fish Tung, Adam W Harley, William Seto, and Katerina Fragkiadaki. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *Proc. ICCV*, volume 2, 2017. 2

[62] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proc. CVPR*, pages 4068–4076, 2015. 3

[63] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proc. CVPR*, 2017. 3

[64] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In *Proc. ICML*, 2017. 5

[65] Mengjiao Wang, Zhixin Shu, Shiyang Cheng, Yannis Panagakis, Dimitris Samaras, and Stefanos Zafeiriou. An adversarial neuro-tensorial approach for learning disentangled

representations. *International Journal of Computer Vision*, 2019. 2

[66] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proc. CVPR*, pages 4724–4732, 2016. 2

[67] Olivia Wiles, A Koepke, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. In *Proc. BMVC*, 2018. 2, 3, 5, 7

[68] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *Proc. ECCV*, 2016. 7

[69] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *Proc. CVPR*, volume 1, 2018. 2

[70] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proc. CVPR*, pages 1385–1392. IEEE, 2011. 2

[71] Xiang Yu, Feng Zhou, and Manmohan Chandraker. Deep deformation network for object landmark localization. In *Proc. ECCV*. Springer, 2016. 7

[72] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection-how to effectively exploit shape and texture features. In *European Conference on Computer Vision*, pages 802–816. Springer, 2008. 7

[73] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2248–2255, 2013. 5

[74] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *Proc. CVPR*, pages 2694–2703, 2018. 1, 2, 3, 5, 6, 7

[75] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *TPAMI*, 38(5):918–930, 2016. 7

[76] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 386–391, 2013. 6

[77] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. CVPR*, 2018. 2, 3, 5, 8

[78] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *Proc. CVPR*, 2015. 7