Learning Human Pose from Unaligned Data through Image Translation

Tomas Jakab
VGG, 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
VGG, University of Oxford
Facebook AI Resesarch, London
vedaldi@robots.ox.ac.uk

Abstract

We introduce a method for learning 2D landmark detectors for human pose from unlabelled video frames and unpaired human keypoints. We formulate this as an image-to-image translation task, similar to CycleGAN, between appearance (RGB) images and skeleton images representing pose. We show CycleGAN confounds appearance and pose information resulting in sub-optimal landmark detection performance. To address this, we introduce a tight keypoint bottleneck which prevents leaking appearance information through skeleton pose images. To facilitate image reconstruction from the clean skeleton images so obtained, we condition the generator on an appearance image sampled from another frame in the video. We further show that the second cycle-consistency constraint in CycleGAN is detrimental to landmark detection performance; we discard it, simplifying the model significantly. Our model with the above components outperforms state-of-the-art unsupervised landmark detection methods on multiple 2D human pose benchmarks. Project page: http://www.robots.ox.ac.uk/~vgg/research/unsupervised_pose/

1. Introduction

A key objective of current research in computer vision is to minimise the dependence on large annotated datasets. We introduce a method for recognizing human pose (2D keypoints) given only videos of humans and *unpaired* keypoint / pose annotations harvested from independent sources, *e.g.* motion capture. We formulate the keypoint detection task as one of image translation: mapping an image of the object to an image of its keypoints / landmarks represented as a skeleton (see fig. 3).

Unpaired image translation can then be tackled by CycleGAN [29]. However, a limitation of CycleGAN is that the mappings between domains are assumed to be one-to-one, which is not the case for most image labelling tasks. While there is a single plausible mapping from an image of a person to its skeleton, the same skeleton can be mapped to many appearance images. CycleGAN removes this ambiguity by introducing in the intermediate skeleton image subtle artefacts that leak appearance information along with pose. This phenomenon, discussed in [5] and visualised in fig. 2, hinders the accuracy of the pose detector (see table 3).

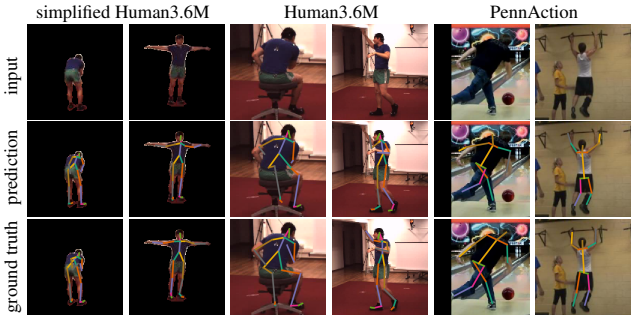

Figure 1. **Learning human pose from unaligned labels.** 2D keypoint predictions (visualised as connected limbs) on the simplified [28] and full Human3.6M [9], and PennAction [27] test sets. Our method learns from unlabelled videos and *unaligned* keypoint annotations, and detects keypoints in complex poses.

We propose several key innovations to address this (section 2). The resulting approach outperforms state-of-the-art unsupervised landmark detection methods on benchmarks such as Human3.6M, without using any paired annotations (section 3).

2. Method

Our aim is to learn a function that maps an image of a person $\boldsymbol{x} \in \mathcal{X} = \mathbb{R}^{3\times H\times W}$ to their pose parameters $\boldsymbol{p}$, represented as K 2D keypoints $\boldsymbol{p} = (p_1, \dots, p_K) \in \Omega^K = \mathbb{R}^{2\times K}$. We learn this function given only pairs of example images $\{(\boldsymbol{x}_i, \boldsymbol{x}'_i)\}_{i=1}^N$ such as frames in a video and unpaired pose parameters $\{\boldsymbol{p}_j\}_{j=1}^M$. We formulate this as an *unsupervised image-to-image translation* task [14, 15, 29], between appearance (RGB) images ($\mathcal{X}$) and *skeleton images* ($\mathcal{Y}$) rendered from the 2D keypoints.

CycleGAN [29] achieves unsupervised image translation by learning mappings which translate between the two domains: $\Phi : \mathcal{X} \to \mathcal{Y}$, and its inverse $\Psi : \mathcal{Y} \to \mathcal{X}$. They ensure that the predictions match true data distributions by aligning them using adversarial discriminators [7] $D_{\mathcal{Y}}$ and $D_{\mathcal{X}}$ respectively. They further enforce *cycle-consistency*

Figure 2. **Appearance leakage.** From left to right: input image, reconstruction, skeleton image, local-contrast normalised (LCN) skeleton image (log scale). To satisfy the cycle-consistency, CycleGAN [29] encodes appearance information in the skeleton image (highlighted in the LCN image).

conditions $\Psi \circ \Phi(\boldsymbol{x}) \approx \boldsymbol{x}$ and $\Phi \circ \Psi(\boldsymbol{y}) \approx \boldsymbol{y}$ to ensure that the mappings are bijective.

As discussed in section 1, the skeleton to image mapping is one-to-many, and enforcing cycle-consistency causes CycleGAN to encode appearance in the skeleton image $\boldsymbol{y}$ [5], leading to sub-optimal pose detection (see fig. 2 and table 3). We overcome this issue through three key innovations (fig. 3):

1. We avoid leaking appearance in the skeleton image by passing it through a tight *bottleneck* implemented using an analytical differentiable renderer (section 2.1).

2. To supplant the lack of appearance information in the clean skeleton images, we condition the generator on an appearance image sampled from another frame in the video (section 2.2).

3. We show that second cycle consistency condition $\Phi \circ \Psi(\boldsymbol{y}) \approx \boldsymbol{y}$ is detrimental to accurate landmark detection (table 3), hence we dispense with it, simplifying the model considerably.

2.1. Skeleton bottleneck

The skeleton bottleneck is implemented as an autoencoder comprising the *skeleton encoder* $\eta : \mathcal{Y} \to \Omega^K$, which maps a skeleton image $\boldsymbol{y}$ to its pose parameters $\boldsymbol{p} = \eta(\boldsymbol{y})$, and the *skeleton generator* $\beta : \Omega^K \to \mathcal{Y}$, which does the opposite.

While the skeleton encoder η is implemented as a ConvNet, crucially, the generator β is *not* learned, but is an analytical and differentiable function that renders the skeleton $\boldsymbol{y}$ from the joint coordinates $\boldsymbol{p}$, thus preventing encoding of appearance in the skeleton image.

Let E be a set of keypoint pairs (i, j) connected by a skeleton edge and let $u \in \{1, \dots, H\} \times \{1, \dots, W\}$ be an image pixel. Then,

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

is the an exponentially-weighted version of the distance transform of the skeleton. The skeleton autoencoder (β, η) is pre-trained using the reconstruction constraint $\eta(\beta(\boldsymbol{p})) = \boldsymbol{p}$ from unpaired pose data and remains fixed as the rest of the model is learned.

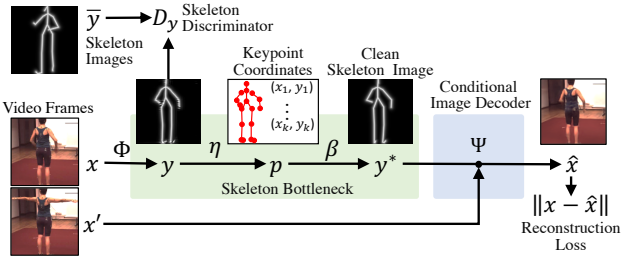

Figure 3. **Model architecture.** Our model learns to detect keypoints $\boldsymbol{p}$ given only video frames $\boldsymbol{x}, \boldsymbol{x}'$ and *unpaired* skeleton images $\bar{\boldsymbol{y}}$. The skeleton bottleneck consists of a pre-trained encoder η and an *analytical* skeleton image renderer β (section 2).

2.2. Conditional image autoencoder

The cycle consistency constraint $\Psi \circ \Phi(\boldsymbol{x}) \approx \boldsymbol{x}$ can be viewed as an autoencoder for the images, where the image code $\Phi(\boldsymbol{x})$ is a skeleton. We retain the encoder $\Phi : \mathcal{X} \to \mathcal{Y}$ as a map from image $\boldsymbol{x}$ to skeleton $\boldsymbol{y}$, but the decoder $\Psi : \mathcal{Y} \times \mathcal{X} \to \mathcal{X}$ now reconstructs the input image $\boldsymbol{x}$ given the skeleton image $\boldsymbol{y}$ as well as a second *conditioning* image $\boldsymbol{x}'$ (*e.g.* another frame from video) supplanting the appearance information missing in the skeleton due to the bottleneck.

The bottleneck of section 2.1 is introduced in the autoencoder $\Psi \circ \Phi$, along with the appearance conditioning to give the following updated cycle consistency constraint:

$$\Psi(\beta \circ \eta \circ \Phi(\boldsymbol{x}), \boldsymbol{x}') \approx \boldsymbol{x}. \tag{2}$$

2.3. Learning objective

Equation (2) is the primary signal for learning our model, and is enforced via a *perceptual loss* [6, 11] $L_{\text{perc}}(\hat{\boldsymbol{x}}, \boldsymbol{x})$ where $\hat{\boldsymbol{x}} = \Psi(\beta \circ \eta \circ \Phi(\boldsymbol{x}), \boldsymbol{x}')$. We also employ an adversarial discriminator $D_{\mathcal{Y}}$ [7] to align the distributions of genuine skeletons $\bar{\boldsymbol{y}}$ and those generated by the image encoder $\boldsymbol{y} = \Phi(\boldsymbol{x})$. We use the squared difference adversarial loss [16]:

$$L_{\text{disc}}(\boldsymbol{y}, \bar{\boldsymbol{y}}) = \mathop{\mathbb{E}}_{\bar{\boldsymbol{y}}\sim\mathcal{Y}}[(1 - D_{\mathcal{Y}}(\bar{\boldsymbol{y}}))^2] + \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{X}}[D_{\mathcal{Y}}(\Phi(\boldsymbol{x}))^2]. \tag{3}$$

A training iteration first samples two frames $(\boldsymbol{x}, \boldsymbol{x}') \sim p(\boldsymbol{x}, \boldsymbol{x}')$ from the same video, and a random skeleton image $\bar{\boldsymbol{y}} \sim p(\boldsymbol{y})$. Next, the autoencoders are evaluated to obtain $\boldsymbol{y} = \Phi(\boldsymbol{x})$, $\boldsymbol{y}^* = \beta(\eta(\boldsymbol{y}))$, and $\hat{\boldsymbol{x}} = \Psi(\boldsymbol{y}^*, \boldsymbol{x}')$. The model is trained to minimise the joint objective $L_{\text{perc}}(\hat{\boldsymbol{x}}, \boldsymbol{x}) + L_{\text{disc}}(\boldsymbol{y}, \bar{\boldsymbol{y}})$. Note, neither constraint (2) nor the discriminator (3) require *labeled images* with pose information.

3. Experiments

We evaluate our method quantitatively for 2D keypoint detection (tables 1 and 2) on simplified and full Human3.6M datasets, and show qualitative results on the PennAction dataset (fig. 1). In table 3, we ablate various components of our model to examine their relative contribution.

	method	all	wait	pose	greet	direct	discuss	walk	eat	phone	purchase	sit	sit down	smoke	take photo	walk dog	walk together
supervised	hourglass [17]	20.22	16.42	14.55	17.58	16.70	20.92	14.11	15.47	19.31	20.45	26.18	40.93	19.68	23.13	22.43	15.41
		19.52	15.53	13.88	17.14	15.81	19.55	**13.74**	15.33	18.81	19.88	25.85	39.07	19.40	22.24	21.58	14.96
	ours	22.30	19.24	21.19	20.23	18.09	22.64	21.69	17.81	22.01	22.32	24.49	28.66	22.83	23.17	26.81	26.36
		19.35	16.12	16.31	16.69	15.66	19.17	13.94	16.41	20.47	19.43	**23.86**	**26.32**	20.96	21.99	21.78	18.57
	ours	23.32	18.81	15.17	17.98	14.90	18.26	21.69	20.83	25.53	19.96	32.98	42.24	23.55	23.57	23.83	25.19
		17.61	**14.16**	**11.88**	**13.78**	**12.62**	**16.44**	17.88	**14.50**	**17.56**	**18.31**	24.05	33.36	**17.65**	**20.10**	**20.42**	**14.01**

Table 1. **2D Human landmark detection.** Comparison on the full Human3.6M test set with supervised baselines — (1) Stacked Hourglass [17], and (2) our model trained with supervision. We report the MSE in pixels for each activity. Shaded rows show error for the original predictions of the model (*no flip* in section 3), while unshaded rows represent the minimum of the errors obtained with and without flipping the predictions against the axis of bilateral symmetry (*with flips* in section 3). The minimum error across all models is in bold.

Datasets. **Human3.6M [9]** is a large-scale motion-capture dataset of accurate poses for human subjects doing 17 different activities, imaged under 4 different viewpoints and a static background. For training, we use subjects 1, 5, 6, 7, and 8, and subjects 9 and 11 for evaluation [23]. **Simplified Human3.6M** [28] contains 6 activities in which human bodies are mostly upright; it comprises 800k training and 90k testing images. **PennAction** [27] contains 2k challenging consumer videos of 15 sports categories.

We split datasets into two *disjoint* parts for sampling image pairs $(\boldsymbol{x}, \boldsymbol{x}')$ (cropped to the provided bounding-boxes), and keypoints $(\bar{\boldsymbol{y}})$ respectively. For Human3.6M datasets, we split the videos in half, while for PennAction, we split in half the set of videos from each action category.

Evaluation. We train our method on each dataset from scratch. For Human3.6m, we follow the standard protocol and report the mean error in pixels for 17 of the 32 joints (but the model estimates all 32 joints). For simplified Human3.6M we follow [28] and report the error for all 32 joints normalized by the image size.

Results. Table 1 summarises results on the full Human3.6M test set. As noted in [10, 28], for unsupervised methods it may be difficult to distinguish the frontal and dorsal views. Following Zhang *et al*. [28] we also report the minimum error between the estimate obtained by the model and its symmetric version (labelled *with flips*). We compare against supervised baselines — (1) the state-of-the-art Stacked Hourglass [17] (using 1-stack), and (2) our model.

method	all	wait	pose	greet	direct	discuss	walk
supervised							
hourglass [17]	**2.16**	**1.88**	**1.92**	**2.15**	**1.62**	**1.88**	**2.21**
unsupervised + supervised linear regression (with flips)							
Thewlis *et al*. [22]	7.51	7.54	8.56	7.26	6.47	7.93	5.40
Zhang *et al*. [28]	4.14	5.01	4.61	4.76	4.45	4.91	4.61
ours *no flips*	7.52	6.86	7.96	7.29	8.34	7.25	7.44
with flips	**2.91**	**2.92**	**2.60**	**2.96**	**2.54**	**2.45**	**3.99**

Table 2. **Simplified 2D human landmark detection.** Comparison with state-of-the-art methods for 2D human landmark detection on the Simplified Human3.6 dataset [28]. We report %-MSE normalised by image size. *no flips / with flips* as in table 1.

With *flips*, our unsupervised model outperforms the baselines overall, however, is worse on challenging activities like sitting and walking. Notably, due to limited variability in the dataset, the supervised baselines overfit on the training set, while our method does not, the training / test errors (*with flips*) are — supervised hourglass: 14.61 / 19.52, ours sup.: 9.67 / 19.35, ours unsup. : 18.00 / 17.61.

Table 2 summarises results on Simplified Human3.6M. Our model significantly outperforms previous methods [22, 28] on all activities, even *without any supervised linear regression* as required by them.

method	humans
CycleGAN	4.39
+ conditional generator	4.07
+ skeleton-bottleneck	3.01
− 2nd cycle = **ours**	2.91

Table 3. **Ablation study.** We start with the CycleGAN [29] model and sequentially augment it with — (1) conditional image generator (Ψ), (2) skeleton bottleneck ($\beta \circ \eta$), and (3) remove the second cycle-consistency constraint (see section 2) resulting in our proposed model. We report 2D landmark detection error (↓ is better) on the simplified Human3.6m (*with flips*; see section 3).

4. Related work

Supervised pose estimation. Estimation of human pose from a single image is explored primarily in the supervised setting, for single [1, 2, 4, 17, 18, 24] and multiple person settings [3, 8]. **Adversarial learning for human pose.** Recent works [12, 26] predict 3D human pose by enforcing re-projection against labelled 2D keypoint images, and a prior from unaligned 3D keypoints using a discriminator. **Unsupervised pose estimation.** Unsupervised 2D pose estimation techniques [13, 19, 21, 22, 25] leverage the relative transformation between different instances of same object. Recent works use image generation to learn unsupervised landmarks, either as a dense deformation field [20, 25], or by learning a neural generator [10, 28]. Our method also relies on image generation, but unlike above methods can map an image to pose without further supervised regression.

5. Conclusion

We presented a method for detecting 2D human pose keypoints, requiring only unlabelled videos and *unaligned* keypoints. We addressed the confounding of appearance and geometry in CycleGAN by introducing a conditioning technique and an analytical and differentiable keypoint bottleneck. Our method outperforms existing unsupervised landmark detection methods and narrows down the gap between supervised and unsupervised methods.

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

References

[1] 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. 3

[2] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *Proc. ECCV*, pages 717–732. Springer, 2016. 3

[3] 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. 3

[4] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Proc. CVPR*, pages 4733–4742, 2016. 3

[5] Casey Chu, Andrey Zhmoginov, and Mark Sandler. Cyclegan, a master of steganography, 2017. 1, 2

[6] 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. 2

[7] 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. 1, 2

[8] 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. 3

[9] 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. 1, 3

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

[11] 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. 2

[12] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proc. CVPR*, 2018. 3

[13] Angjoo Kanazawa, David W Jacobs, and Manmohan Chandraker. Warpnet: Weakly supervised matching for single-view reconstruction. In *Proc. CVPR*, pages 3253–3261, 2016. 3

[14] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017. 1

[15] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016. 1

[16] 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. 2

[17] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proc. ECCV*, 2016. 3

[18] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *Proc. CVPR*, pages 1913–1921, 2015. 3

[19] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proc. CVPR*, volume 2, 2017. 3

[20] Zhixin Shu, Mihir Sahasrabudhe, Riza 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 (ECCV)*, pages 650–665, 2018. 3

[21] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *Proc. NIPS*, 2017. 3

[22] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *Proc. ICCV*, 2017. 3

[23] 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. 3

[24] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proc. CVPR*, pages 4724–4732, 2016. 3

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

[26] 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. 3

[27] 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. 1, 3

[28] 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, 3

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. ICCV*, 2018. 1, 2, 3