Conditional Image Generation for Learning the Structure of Visual Objects

Tomas Jakab[1*] **Ankush Gupta**[1*] **Hakan Bilen**[2] **Andrea Vedaldi**[1]

[1] Visual Geometry Group
University of Oxford
{tomj,ankush,vedaldi}@robots.ox.ac.uk

[2] School of Informatics
University of Edinburgh
hbilen@ed.ac.uk

Abstract

In this paper, we consider the problem of learning landmarks for object categories without any manual annotations. We cast this as the problem of conditionally generating an image of an object from another one, where the images differ by acquisition time and/or viewpoint. The process is aided by providing the generator with a keypoint-like representation extracted from the target image through a tight bottleneck. This encourages the representation to distil information about the object geometry, which changes from source to target, while the appearance, which is shared between the source and target, is read off from the source alone. Conditioning simplifies the generation task significantly, to the point that adopting a simple perceptual loss instead of more sophisticated approaches such as adversarial training is sufficient to learn landmarks. We show that our method is applicable to a large variety of datasets — faces, people, 3D objects, and digits — without any modifications. We further demonstrate that we can learn landmarks from synthetic image deformations or videos, all without manual supervision, while outperforming state-of-the-art unsupervised landmark detectors.

1 Introduction

There is a growing interest in developing learning methods with reduced or no dependence on manual supervision, and minimise the burden of gathering massive amounts of labelled data. In this paper we consider learning the *structure* of visual objects given only unlabelled images. To a first approximation, this can be reduced to the problem of learning a set of object landmarks, such as the nose, the eyes, and the mouth of a face, or the positions of hands, shoulders, and head in a human body. Landmarks capture the geometric structure of objects, and help establish meaningful correspondences between their images.

Our approach learns from pairs of images of objects that differ by time and/or viewpoint. Such pairs may be extracted from a video sequence or generated from synthetic perturbations. The method is based the idea of *conditional image generation*. Namely, we generate a target image conditioned on observing a source image, where they differ in their time of acquisition, or viewpoint. However, it is difficult to unambiguously predict motion from a single image. Hence, we aid the process by supplementing the generator with a compressed representation of the target image. The goal of this representation is to *distil the essential nature of the change*. Since the main factor of variation is the pose of the underlying object, the representation should learn to capture it, while transferring from the source image the appearance or style of the object.

The general idea of using conditional image generation for learning the structure of visual data has already been explored in the context of (variational) auto-encoders, and Generative Adversarial Networks (GAN [12]; see section 2). The key challenge is to encourage the representation to extract

[*] equal contribution.

Preprint. Work in progress.

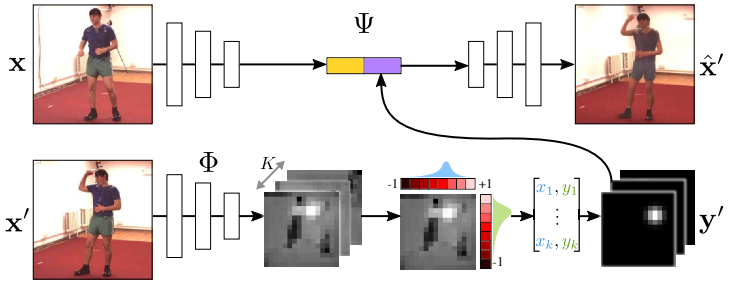


Figure 1: **Model Architecture.** Given a pair of source and target images $(\mathbf{x}, \mathbf{x}')$, the pose-regressor Φ extracts K heatmaps from $\mathbf{x}'$, which are then marginalized to estimate coordinates of keypoints, to limit the information flow. 2D Gaussians $(\mathbf{y}')$ are rendered from these keypoints and stacked along with the image features extracted from $\mathbf{x}$, to reconstruct the target as $\Psi(\mathbf{x}, \mathbf{y}') = \hat{\mathbf{x}}'$. By restricting the information-flow our model learns semantically meaningful keypoints, without any annotations.

only pose-related information, and do so in a form which can be readily ingested by the generator. Our contribution is to propose a *representation bottleneck* which encourages the extraction of information highly-related to object landmarks (section 3). The strength of the method resides in the simplicity of the setup and its ability to work well without modifications on far more complex data than that amenable to previous unsupervised landmark learning methods. Furthermore, we show that the method can learn from synthetically-generated image deformations or raw videos directly as it *does not* require access to information about correspondences, optical-flow, or transformation, unlike other methods [43, 44].

The key idea is to define the compressed representation to be similar to a collection of heatmaps of detected landmarks. We do so by squeezing heatmaps into single 2D point-estimates of each landmark's location, and then expand them back to clean Gaussian heatmaps for further processing by a generator. Inspired by the work of Chen *et al*. [4], we then generate the target image not by warping the source image, but by learning a conditional image generator network that takes as input the source image and the pseudo-landmark heatmaps. We show that the two ideas of conditional image generation and shaping of the representation space can indeed learn good proxies to object landmarks in a completely unsupervised manner, separating object style from geometry.

A critical advantage of conditional image generation is that it simplifies the generation task considerably, making it much easier to learn a generator network [17]. While, the default choice here would have been to use techniques like GANs, we take the more direct approach of using a perceptual loss as in [8], and show that this not only obtains excellent generation results, but more importantly, discovers meaningful object structure (section 4).

2 Related work

The recent approaches of [43, 44] can learn to extract landmarks based on the principles of equivariance and distinctiveness. In contrast to our work, these methods are not generative and only focus on accurate localisation of landmarks. Further, they rely on known correspondences between images, which are obtained either through optical flow, or synthetic transformations, and hence, cannot leverage video data. Since the principle of equivariance is orthogonal to our approach it can be incorporated as an additional cue in our method.

Unsupervised learning of representations has traditionally been achieved using autoencoders and restricted Boltzmann machines [14, 46, 13]. Word2Vec [28] learns skip-thought vectors [22] with compositional properties, also demonstrated on images by [37] using GANs. InfoGAN [5] uses GANs to disentangle factors in the data by imposing a certain structure in the latent space. Our approach also works by imposing a latent structure, but using a *conditional* encoder instead of an auto-encoder.

Learning representations using conditional image generation via a bottleneck was demonstrated by Xue *et al*. [50] in variational autoencoders [21], and by Whitney *et al*. [49] using a discrete gating mechanism to combine representations of successive video frames. Denton *et al*. [7] factor the pose and identity in videos through an adversarial loss on the pose embeddings. We instead design our

bottleneck explicitly, shaping features to resemble the output of a landmark detector, without any adversarial training. Villegas *et al*. [45] also generate future frames by extracting a representation of appearance and human pose, but, differently from us, they require ground-truth pose annotations. Several other generative approaches [42, 40, 38, 47, 33] focus on video extrapolation. Srivastava *et al*. [40] employ Long Short Term Memory (LSTM) [15] networks to encode video sequences into fixed length representation and decode it to reconstruct the input sequence. Vondrick *et al*. [47] propose a GAN for videos, also with a spatio-temporal convolutional architecture that disentangles foreground and background to generate realistic frames. The Video Pixel Networks [19] estimate the discrete joint distribution of the pixel values in a video by encoding different modalities such as time, space and colour information. In contrast, we learn a *structured embedding* that explicitly encodes image landmarks.

Finally, the concurrent work by [53] shares several similarities with ours, in that they also use conditional image generation with the goal of learning landmarks. However, there are key differences in how these ideas are applied. In particular, their method is based on generating a single image from itself using landmark-transported features. This, as we show in the experiments, is insufficient to learn geometry and requires, as they do, to also incorporate the principle of equivariance [43]. This is a key difference with our method, as ours results in a much simpler system that does *not* require to know the optical flow / correspondences between images, and can learn from raw videos directly.

3 Method

Let $\mathbf{x}, \mathbf{x}' \in \mathcal{X} = \mathbb{R}^{H\times W\times C}$ be two images of an object, for example extracted as frames in a video sequence, or synthetically generated by randomly deforming $\mathbf{x}$ into $\mathbf{x}'$. Conventionally, we call $\mathbf{x}$ the source image and $\mathbf{x}'$ the target image and we denote the image domain, namely the $H\times W$ lattice, Ω.

We are interested in learning a function $\Phi(\mathbf{x}) = \mathbf{y} \in \mathcal{Y}$ that captures the "structure" of the object in the image, which we cast as the problem of detecting K object landmarks. As a first approximation, although we are going to modify this later, assume then that $\mathbf{y} = (u_1, \dots, u_K) \in \Omega^K = \mathcal{Y}$ are K coordinates $u_k \in \Omega$, one per landmark.

In order to learn the map Φ in an unsupervised manner, we consider the problem of conditional image generation. Namely, we wish to learn a generator function

$$\Psi : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}, \qquad (\mathbf{x}, \mathbf{y}') \mapsto \mathbf{x}'$$

such that the target image $\mathbf{x}' = \Psi(\mathbf{x}, \Phi(\mathbf{x}'))$ is reconstructed from the *source image* $\mathbf{x}$ and the *representation* $\mathbf{y}' = \Phi(\mathbf{x}')$ of the *target image*. In practice, we learn both functions Φ and Ψ jointly to minimise the expected reconstruction loss $\min_{\Psi,\Phi} E_{\mathbf{x},\mathbf{x}'}\left[\mathcal{L}(\mathbf{x}', \Psi(\mathbf{x}, \Phi(\mathbf{x}')))\right]$. Note that, if we do not restrict the form of $\mathcal{Y}$, then a trivial solution to this problem is to learn identity mappings by setting $\mathbf{y}' = \Phi(\mathbf{x}') = \mathbf{x}'$ and $\Psi(\mathbf{x}, \mathbf{y}') = \mathbf{y}'$. However, given that $\mathbf{y}'$ has the "form" of a set of landmark detections, the model is strongly encouraged to learn those. This is explained next.

3.1 Implementation with heatmaps

In order for the model $\Phi(\mathbf{x})$ to learn to extract keypoint-like structures from the image, we terminate the network Φ with a layer that forces the output to be akin to a set of K keypoint detections. This is done in three steps. First, K heatmaps $S_u(\mathbf{x}; k), u \in \Omega$ are generated, one for each keypoint $k = 1, \dots, K$. These heatmaps are obtained in parallel as the channels of a $\mathbb{R}^{H\times W\times K}$ tensor using a standard convolutional neural network architecture. Second, each heatmap is renormalised to a probability distribution and condensed to a point by computing the (spatial) expected value of the latter:

$$u_k^*(\mathbf{x}) = \frac{\sum_{u\in\Omega} u e^{S_u(\mathbf{x};k)}}{\sum_{u\in\Omega} e^{S_u(\mathbf{x};k)}} \tag{1}$$

Third, each heatmap is replaced with a Gaussian-like function centred at u_k^* with a small fixed standard deviation:

$$\Phi_u(\mathbf{x}; k) = \exp\left(-\frac{1}{2\sigma^2}\|u - u_k^*(\mathbf{x})\|^2\right) \tag{2}$$

The end result is a new tensor $\mathbf{y} = \Phi(\mathbf{x}) \in \mathbb{R}^{H\times W\times K}$ with the location of K maxima. Since it is possible to recover the landmark locations exactly from these heatmaps, this representation is

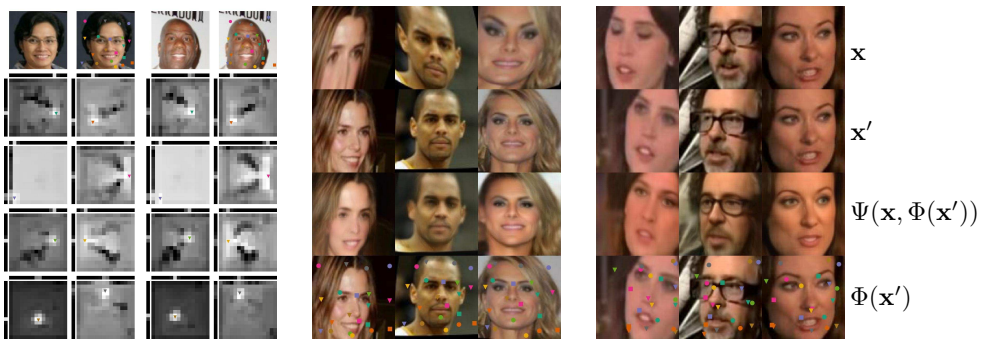


Figure 2: **Unsupervised Landmarks. [left]:** Two example images with selected landmarks u_k overlaid and their corresponding 2D score maps $S_u(\mathbf{x}; k)$ (see section 3.1; brighter pixels indicate higher confidence). The distribution of background landmarks is spread (*e.g.* row-2, col-1) whereas for foreground ones it is peaked. **[middle]:** CelebA images showing the synthetically transformed source and target images $\mathbf{x}$ and $\mathbf{x}'$ images, the reconstructed target $\Psi(\mathbf{x}, \Phi(\mathbf{x}'))$, and the unsupervised landmarks $\Phi(\mathbf{x}')$. **[right]:** The same for video frames from VoxCeleb.

equivalent to the one considered above (2D coordinates); however, it is more useful as an input to a generator network, as discussed later.

One may wonder whether this construction can be simplified by removing steps two and three and simply consider $S(\mathbf{x})$ (possibly after re-normalisation) as the output of the encoder $\Phi(\mathbf{x})$. The answer is that these steps, and especially eq. (1), ensure that very little information from $\mathbf{x}$ is retained, which, as suggested above, is key to avoid degenerate solutions. Converting back to Gaussian landmarks in eq. (2), instead of just retaining 2D coordinates, ensures that the representation is still utilisable by the generator network.

Separable implementation. In practice, we consider a separable variant of eq. (1) for computational efficiency. Namely, let $u = (u_1, u_2)$ be the two components of each pixel coordinate and write $\Omega = \Omega_1 \times \Omega_2$. Then we set

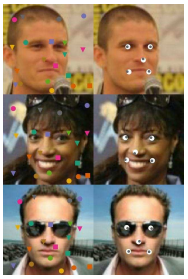
$$u^*_{ik}(\mathbf{x}) = \frac{\sum_{u_i \in \Omega_i} u_i e^{S_{u_i}(\mathbf{x};k)}}{\sum_{u_i \in \Omega_i} e^{S_{u_i}(\mathbf{x};k)}}, \qquad S_{u_i}(\mathbf{x}; k) = \sum_{u_j \in \Omega_j} S_{(u_1,u_2)}(\mathbf{x}; k),$$

where $i = 1, 2$ and $j = 2, 1$ respectively. Figure 2 visualizes the source $\mathbf{x}$, target $\mathbf{x}'$, generated $\Psi(\mathbf{x}, \Phi(\mathbf{x}'))$, and the unsupervised-landmarks $\Phi(\mathbf{x}')$ on the target and generated images, along with the heatmaps $S_u(\mathbf{x}; k)$ and marginalized separable soft-max distributions on the top and left of each heatmap, for $K = 20$ keypoints and two samples from the CelebA [24] training set.

3.2 Generator network using a perceptual loss

The goal of the generator network $\hat{\mathbf{x}}' = \Psi(\mathbf{x}, \mathbf{y}')$ is to map the source image $\mathbf{x}$ and the distilled version $\mathbf{y}'$ of the target image $\mathbf{x}'$ to a reconstruction of the latter. Thus the generator network is optimised to minimise a reconstruction error $\mathcal{L}(\mathbf{x}', \hat{\mathbf{x}}')$.

The design of the reconstruction error is important for good performance. Nowadays the standard practice is to learn such a loss function using adversarial techniques, as exemplified in numerous variants of GANs. However, since the goal here is not reconstruction per se, but rather to induce a representation $\mathbf{y}'$ of the object geometry, a simpler method may suffice. Inspired by the recent excellent results for photo-realistic image synthesis of [4], we resort here to use the "content representation" or "perceptual" loss used successfully for various generative networks [11, 1, 9, 18, 27, 31, 32]. The perceptual loss compares a set of the activations extracted from multiple layers of a deep network for both the reference and the generated images, instead of the only raw pixel values. We define the loss as $\mathcal{L}(\mathbf{x}', \hat{\mathbf{x}}') = \sum_l \alpha_l \|\Gamma_l(\mathbf{x}') - \Gamma_l(\hat{\mathbf{x}}')\|_2^2$, where $\Gamma(\mathbf{x})$ is an off-the-shelf pre-trained neural network, for example VGG-19 [39], Γ_l denotes the output of the l-th subnetwork (obtained by chopping Γ at layer l). As our goal is to have a purely-unsupervised learning, we pre-train the network by using a self-supervised approach — colorising grayscale images [25]. We also test using a VGG-19 model pre-trained for image classification in ImageNet. All other networks are trained from scratch.



n supervised	Thewlis [43] warp	Ours warp	Ours no warp	Ours selfsup warp	Ours selfsup no warp
1	10.82	22.83 ± 9.55	12.35 ± 2.05	10.84 ± 3.22	10.95 ± 2.22
5	9.25	8.06 ± 0.83	15.25 ± 1.36	7.35 ± 1.26	15.97 ± 4.13
10	8.49	7.02 ± 0.73	14.27 ± 3.89	6.23 ± 0.17	14.00 ± 1.66
100	—	4.99 ± 0.19	5.20 ± 0.32	4.72 ± 0.10	5.11 ± 0.06
500	—	4.48 ± 0.06	3.60 ± 0.05	4.19 ± 0.02	3.42 ± 0.07
1000	—	4.36 ± 0.07	3.38 ± 0.04	4.12 ± 0.02	3.23 ± 0.01
All (19,000)	7.15	4.24 ± 0.02	$3.23\pm$ N/A	4.02 ± 0.01	$3.08\pm$ N/A

Figure 3: **Supervised Regression on MAFL. [left]:** Supervised linear regression of 5 keypoints (right) from 20 unsupervised (left) on MAFL test set. Centre of the white-dots correspond to the ground-truth location, while the dark ones are the predictions. **[right]:** MSE ($\pm\sigma$) on the MAFL test-set for varying number (n) of supervised samples from MAFL training set used for learning the regressor from 30 unsupervised landmarks.

The parameters $\alpha_l > 0, l = 1, \ldots, n$ are scalars that balance the terms. We use a linear combination of the reconstruction error for 'input', 'conv1_2', 'conv2_2', 'conv3_2', 'conv4_2' and 'conv5_2' layers of VGG-19; $\{\alpha_l\}$ are updated online during training to normalise the expected contribution from each layer as in [4]. However, we use the ℓ_2 norm instead of their ℓ_1, as it worked better for us.

4 Experiments

This section assesses our method extensively. We start by providing the details of the landmark detection and generator networks in section 4.1; a common architecture is used across all the experiments. Then we evaluate the accuracy of the detection network on facial (section 4.2), human-body (section 4.3) landmark localisation, both qualitatively and quantitatively. In section 4.4 we analyse the invariance of the learned landmarks to various nuisance factors, and finally in section 4.5, study the factorised representation of object style and geometry in the generator.

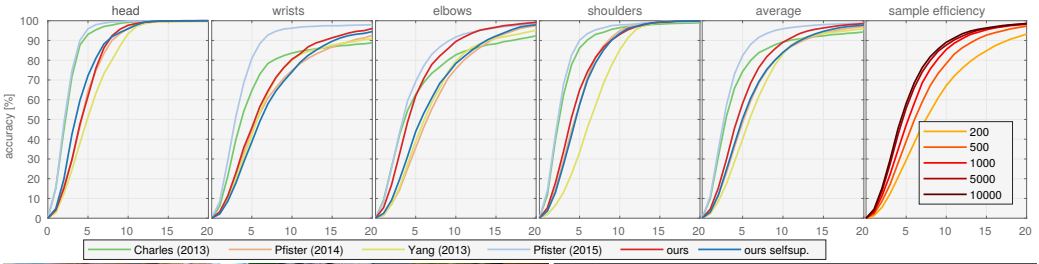
4.1 Model details

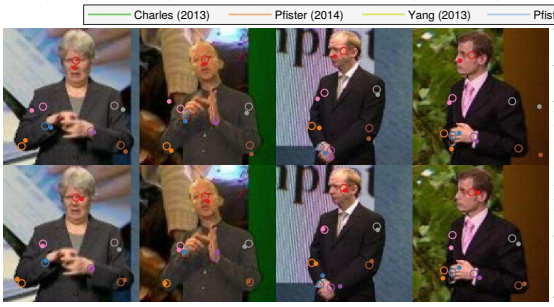
Landmark detection network. The landmark detector ingests the image $\mathbf{x}'$ to produce K landmark heatmaps $\mathbf{y}'$. It is composed of sequential blocks consisting of two convolutional layers each. All the layers use 3×3 filters, except the first one which uses 7×7. Each block doubles the number of feature-channels in the previous block, with 32 channels in the first one. The first layer in each block, except the first block, downsamples the input tensor using stride-2 convolution. The spatial size of the final output, outputting the heatmaps, is set to 16×16. Thus, due to downsampling, for a network with $n-3$, $n \geq 4$ blocks, the resolution of the input image is $H\times W = 2^n\times2^n$, resulting in $16\times16\times(32\cdot2^{n-3})$ tensor. A final 1×1 convolutional layer maps this tensor to a final $16\times16\times K$ tensor, with one layer per landmark. As described in section 3.1, these K feature-channels are then used to render $16\times16\times K$ 2D-Gaussian maps $\mathbf{y}'$ (with $\sigma = 0.1$).

Image generation network. The image generator takes as input the image $\mathbf{x}$ and the landmarks $\mathbf{y}' = \Phi(\mathbf{x}')$ extracted from the second image in order to reconstruct the latter. This is achieved in two steps: first, the image $\mathbf{x}$ is encoded as a feature tensor $\mathbf{z} \in \mathbb{R}^{16\times16\times C}$ using a convolutional network with exactly the same architecture as the landmark detection network except for the final 1×1 convolutional layer, which is omitted; next, the features $\mathbf{z}$ and the landmarks $\mathbf{y}'$ are stacked together (along the channel dimension) and fed to a regressor that reconstructs the target frame $\mathbf{x}'$.

The regressor also comprises of sequential blocks with two convolutional layers each. The input to each successive block, except the first one, is upsampled two times through bilinear interpolation, while the number of feature channels is halved; the first block starts with 256 channels, and a minimum of 32 channels are maintained till a tensor with the same spatial dimensions as $\mathbf{x}'$ is obtained. The final layer regresses the three RGB channels with no further non-linearity. All layers use 3×3 filters and each block has two layers similarly to the landmark network.

All the weights are initialised with random gaussian noise ($\sigma = 0.01$), and optimised using Adam [20] with a weight decay of $5\cdot10^{-4}$. The learning rate is set to 10^{-2}, and lowered by a factor of 10 once the training error stops decreasing; the ℓ_2-norm of the gradients is bounded to 1.0.





BBC Pose Accuracy (%) at d = 6 pixels

	Head	Wrsts	Elbws	Shldrs	Avg.
Pfister *et al*. [35]	98.00	88.45	77.10	93.50	88.01
Charles *et al*. [3]	95.40	72.95	68.70	90.30	79.90
Chen *et al*. [6]	65.9	47.9	66.5	76.8	64.1
Pfister *et al*. [36]	74.90	53.05	46.00	71.40	59.40
Yang *et al*. [51]	63.40	53.70	49.20	46.10	51.63
Ours (selfsup.)	81.10	49.05	53.05	70.10	60.79
Ours	76.10	56.50	70.70	74.30	68.44

Figure 4: **Learning Human Pose.** 50 unsupervised keypoints are learnt on the BBC Pose dataset. Annotations (empty circles in the images) for 7 keypoints are provided, corresponding to — head, wrists, elbows and shoulders. Solid circles represent the predicted positions; in **[fig-top]** these are raw discovered keypoints which correspond maximally to each annotation; in **[fig-bottom]** these are regressed (linearly) from the discovered keypoints. **[table]:** Comparison against *supervised* methods; %-age of points within d= 6-pixels of ground-truth is reported. **[top-row]:** accuracy-vs-distance d, for each body-part; **[top-row-rightmost]:** average accuracy for varying number of supervised samples used for regression.

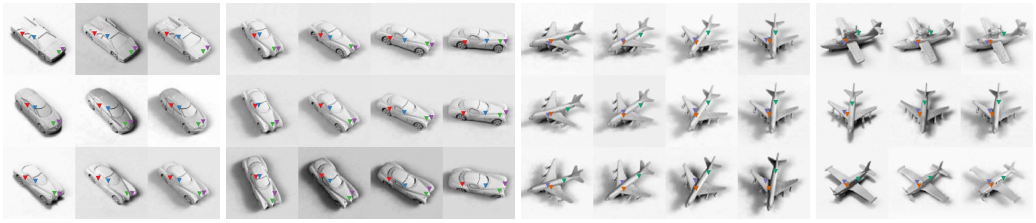
4.2 Learning facial landmarks

Setup. We explore extracting source-target image pairs $(\mathbf{x}, \mathbf{x}')$ using either (1) synthetic transformations, or (2) videos. In the first case, the pairs are obtained as $(\mathbf{x}, \mathbf{x}') = (g_1\mathbf{x}_0, g_2\mathbf{x}_0)$ by applying two random thin-plate-spline transformations (TPS) [10, 48] g_1, g_2 to sample images $\mathbf{x}_0$. We use the 200k CelebA [24] images after resizing them to 128×128 resolution. The dataset provides annotations for 5 facial landmarks — eyes, nose and mouth corners, which we *do not* use for training. Following [43] we exclude the images in MAFL [55] test-set from the training split and generate synthetically-deformed pairs as in [43, 53], but the transformations themselves are not required for training.

In the second case, $(\mathbf{x}, \mathbf{x}')$ are two frames sampled from a video. We consider VoxCeleb [29], a large dataset of face tracks, consisting of 1251 celebrities speaking over 100k English language words. We use the standard training split and remove any overlapping identities which appear in the test sets. Pairs of frames from the same video, but possibly belonging to different word utterances, are randomly sampled for training. By using video data for training our models we eliminate the need for engineering synthetic data.

Qualitative results. Figure 2 shows the learned heatmaps and source-target-reconstruction-keypoints quadruplets $\langle \mathbf{x}, \mathbf{x}', \Psi\left(\mathbf{x}, \Phi(\mathbf{x}')\right), \Phi(\mathbf{x}')\rangle$ for synthetic transformations and videos. We note that the method reconstructs accurately the target image using the extracted keypoints, which consistently track facial features across deformation and identity changes (*e.g.*, the green circle roughly tracks the lower chin and the light blue square is between the eyes). The regressed semantic keypoints on the MAFL test set are visualised in fig. 3, where they are localised with high accuracy.

Figure 5: **Unsupervised Landmarks on Human3.6M. [left]:** an example quadruplet source-target-reconstruction-keypoint (left to right) from Human3.6M. **[right]:** learned keypoints on a test video sequence. The landmarks consistently track the legs, arms, torso and head across frames.



(a) (b) (c) (d)

Figure 6: **Invariant Localisation.** Unsupervised keypoints discovered on smallNORB test set for the *car* and *airplane* categories. Out of 20 learned keypoints, we show the most geometrically stable ones: they are invariant to pose, shape, and illumination. **[b–c]:** elevation-vs-azimuth; **[a, d]:** shape-vs-illumination (y-axis-vs-x-axis).

Quantitative results. We follow [43, 44] and use unsupervised keypoints learnt on CelebA and VoxCeleb to regress manually-annotated keypoints in the MAFL and AFWL [23] test sets. We freeze the parameters of the unsupervised detector network (Φ) and learn a *linear* regressor (without bias) from our unsupervised keypoints to 5 manually-labelled ones.

We report results in terms of standard MSE (normalised by the inter-ocular distance [55]) and show a few regressed keypoints in fig. 3. The data is optionally augmented using random TPS transformations as in [43, 44], referred to as "warp"/"no-warp" in the table.

Sample efficiency. Figure 3 reports the performance of detectors trained on CelebA as a function of the number n of supervised examples used to translate from unsupervised to supervised keypoints. We note that $n = 5$ is already sufficient for results comparable to the previous state-of-the-art (SoA) and that performance saturates at $n \approx 100$ (vs. 19,000 available training samples). Warp augmentation is useful only for low values of n and harmful otherwise and the self-supervised perceptual loss is better than the supervised one.

Vs. SoA. Table 1 compares our regression results to the SoA. We experiment regressing from $K{=}\{10, 30, 50\}$ unsupervised landmarks, "no-warp" setting, and using the self-supervised or the supervised perceptual loss networks; the number of samples n used for regression is maxed out to be consistent with the reported results. On MAFL, at (3.08%) error we significantly outperform the unsupervised methods of [44] (6.67%) and are slightly better than the concurrent work by [53] (3.16%). On AFWL, [53] is slightly better $(6.98\%$ vs $6.58\%)$. Compared to the latter, however, our method is much simpler and does not require synthetic warps for learning. When synthetic warps are removed [53], so that the *equivariance constraint cannot be employed*, our method is far better $(3.08\%$ vs $8.42\%)$. We are also significantly better than many SoA *supervised* detectors [52, 41, 55] (even when n is reduced to 100 examples). Finally, training with VoxCeleb video frames slightly degrades the performance on MAFL due to domain gap, but improves on AFLW.

Method	K	MAFL	AFLW
RCPR [2]		–	11.60
CFAN [52]		15.84	10.94
TCDCN [55]		7.95	7.65
Cascaded CNN [41]		9.73	8.97
RAR [41]		–	7.23
MTCNN [54]		5.39	6.90
Thewlis [43]	50	6.67	10.53
Thewlis [44](frames)	–	5.83	8.80
Zhang [53]	10	3.46	7.01
w/ equiv.	30	3.16	**6.58**
w/o equiv.	30	8.42	–
Ours, training set: CelebA			
loss-net: sup.	10	4.89	8.78
	30	3.23	7.20
	50	3.33	7.34
loss-net: selfsup.	10	4.69	8.28
	30	**3.08**	6.98
	50	3.90	7.89
Ours, training set: VoxCeleb			
loss-net: sup.	30	4.17	7.10

Table 1: **Comparison with state-of-the-art on MAFL and AFLW.** K is the number of unsupervised landmarks.

4.3 Learning human body landmarks

Setup. Articulated limbs make landmark localisation on human body significantly more challenging than faces. We consider two *video* datasets, BBC-Pose [3], and Human3.6M [16]. BBC-Pose comprises of 20 one-hour long videos of sign-language signers with varied appearance, and dynamic background; the test set includes 1000 frames. The frames are annotated with 7 keypoints corresponding to head, wrists, elbows, and shoulders which, as for faces, we use only for quantitative evaluation, not for training. Human3.6M dataset contains videos of 11 actors in various poses, shot

Figure 7: **Disentangling style and geometry.** Image generation conditioned on *spatial* keypoints induces disentanglement of representations for style and geometry in the generator. Source image ($\mathbf{x}$) imparts style (*e.g.* colour, texture), while the target image ($\mathbf{x}'$) influences the geometry (*e.g.* shape, pose). Here, during inference, $\mathbf{x}$ **[middle]** is sampled to have a different *style* than $\mathbf{x}'$ **[top]**, although during training, image pairs with *consistent* style were sampled. The generated images **[bottom]** borrow their style from $\mathbf{x}$, and geometry from $\mathbf{x}'$. **(a) SVHN Digits:** the foreground and background colours are swapped. **(b) AFLW Faces:** pose of the style image $\mathbf{x}$ is made consistent with $\mathbf{x}'$. **(c) Human3.6M:** the background, hat, and shoes are retained from $\mathbf{x}$, while the pose is borrowed from $\mathbf{x}'$. All images are sampled from respective test sets, never seen during training.

from multiple viewpoints. Image-pairs are extracted by randomly sampling frames from the the same video sequence, with the additional constraint of maintaining the time difference within the range 3-30 frames for Human3.6M. Loose crops around the subjects are extracted using the provided annotations and resized to 128×128 pixels. Detectors for $K = 50$ keypoints are trained.

Qualitative results. Figure 4 shows raw unsupervised keypoints and the regressed semantic ones on the BBC-Pose dataset. For each annotated keypoint, a maximally matching unsupervised keypoint is identified, by solving bipartite linear assignment, using mean distance as the cost. Regressed keypoints consistently track the annotated points. Figure 5 shows $\langle \mathbf{x}, \mathbf{x}', \Psi\left(\mathbf{x}, \Phi(\mathbf{x}')\right)\rangle$ quadruplets, as for faces, as well as the discovered keypoints. All the keypoints lie on top of the human actors, and consistently track the body across identities and poses. However, the model cannot discern frontal and dorsal sides of the human body apart, possibly due to weak cues in the images, and no explicit constraints enforcing such consistency.

Quantitative results. Figure 4 compares the accuracy of localising the 7 keypoints on BBC-Pose against *supervised* methods, for both self-supervised and supervised perceptual loss networks. The accuracy is computed as the the $\%$-age of points within a specified pixel distance d. In this case, the top two supervised methods are better than our unsupervised approach, but we outperform [34, 51] using 1k training samples (vs. 10k); furthermore, methods such as [35] are specialised for videos and leverage temporal smoothness. Training using the supervised perceptual loss is understandably better than using the self-supervised one. Performance is particularly good on parts such as the elbow.

4.4 Learning 3D object landmarks: pose, shape, and illumination invariance

We train our unsupervised keypoint detectors on the SmallNORB [26] dataset, comprising 5 object categories with 10 object instances each, imaged from regularly spaced viewpoints and under different illumination conditions. We train category-specific detectors for $K = 20$ keypoints using image-pairs from neighbouring viewpoints and show results in fig. 6 for *car* and *airplane*. Keypoints most invariant to various factors are visualised. These landmarks are especially robust to changes in illumination and elevation angle. They are also invariant to smaller changes in azimuth ($\pm80^\circ$), but fail to generalise beyond that. Most interesting, they localise structurally similar regions, even when there is a large change in object shape (*e.g.* fig. 6-(d)); such landmarks could thus be leveraged for viewpoint-invariant semantic matching.

4.5 Disentangling appearance and geometry

In fig. 7 we show that our method can be interpreted as disentangling appearance from geometry. Generator/ keypoint networks are trained on SVHN digits [30], ALFW faces, and Human3.6M people. The generator network is capable of retaining the geometry of an image, and substituting the style with any other image in the dataset, including unrelated image pairs never seen during training. For example, in the third column we re-render the number 3 by mixing its geometry with the appearance of the number 5. This generalises significantly from the training examples, which only consist of pairs of digits sampled from the *same* house-number instance, sharing a common style.

5 Conclusions

In this paper we have shown that a simple network trained for conditional image generation can be utilised to induce, without manual supervision, a set of object landmarks. We have demonstrated this on faces, obtaining competitive results with previous unsupervised as well as supervised methods for landmark detection. Importantly, we have also shown that the method can extend to much more challenging cases, such as detecting landmarks of people, 3D objects and digits.

Acknowledgments. We are grateful for the support provided by EPSRC AIMS CDT, ERC-IDIU, and the Clarendon Fund scholarship. We would like to thank James Thewlis for suggestions and support with code and data.

References

[1] J. Bruna, P. Sprechmann, and Y. LeCun. Super-resolution with deep convolutional sufficient statistics. In *Proc. ICLR*, 2016.

[2] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1513–1520. IEEE, 2013.

[3] J. Charles, T. Pfister, D. Magee, D. Hogg, and A. Zisserman. Domain adaptation for upper body pose tracking in signed TV broadcasts. In *Proc. BMVC*, 2013.

[4] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *Proc. ICCV*, volume 1, 2017.

[5] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proc. NIPS*, pages 2172–2180, 2016.

[6] X. Chen and A. L. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Proc. NIPS*, 2014.

[7] E. L. Denton and V. Birodkar. Unsupervised learning of disentangled representations from video. In *Proc. NIPS*. 2017.

[8] A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. In *Proc. NIPS*, pages 658–666, 2016.

[9] A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. In *Proc. NIPS*, 2016.

[10] J. Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive theory of functions of several variables*. 1977.

[11] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proc. CVPR*, 2016.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proc. NIPS*, 2014.

[13] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[14] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[15] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[16] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014.

[17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. CVPR*, 2017.

[18] J. Johnson, A. Alahi, and F. Li. Perceptual losses for real-time style transfer and super-resolution. In *Proc. ECCV*, 2016.

[19] N. Kalchbrenner, A Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu. Video pixel networks. *arXiv preprint arXiv:1610.00527*, 2016.

[20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[22] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *Proc. NIPS*, pages 3294–3302. 2015.

[23] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCV Workshops*, 2011.

[24] Ziwei L., Ping L., Xiaogang W., and Xiaoou T. Deep learning face attributes in the wild. In *Proc. ICCV*, 2015.

[25] G. Larsson, M. Maire, and G. Shakhnarovich. Learning representations for automatic colorization. In *Proc. ECCV*, 2016.

[26] Y. LeCun, F. J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proc. CVPR*, 2004.

[27] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. CVPR*, 2017.

[28] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[29] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017.

[30] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS DLW*, volume 2011, 2011.

[31] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Proc. NIPS*, 2016.

[32] A. Nguyen, J. Yosinski, Y. Bengio, A. Dosovitskiy, and J. Clune. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proc. CVPR*, 2017.

[33] V. Patraucean, A. Handa, and R. Cipolla. Spatio-temporal video autoencoder with differentiable memory. In *ICLR Workshop*, 2015.

[34] T. Pfister, J. Charles, and A. Zisserman. Large-scale learning of sign language by watching TV (using co-occurrences). In *Proc. BMVC*, 2013.

[35] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *Proc. ICCV*, 2015.

[36] T. Pfister, K. Simonyan, J. Charles, and A. Zisserman. Deep convolutional neural networks for efficient pose estimation in gesture videos. In *Proceedings of the Asian Conference on Computer Vision*, 2014.

[37] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proc. ICLR*, 2016.

[38] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In *Proc. NIPS*, pages 217–225, 2016.

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[40] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using lstms. In *Proc. ICML*, pages 843–852, 2015.

[41] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proc. CVPR*, 2013.

[42] I. Sutskever, G. E. Hinton, and G. W. Taylor. The recurrent temporal restricted boltzmann machine. In *Proc. NIPS*, pages 1601–1608, 2009.

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

[44] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised object learning from dense invariant image labelling. In *Proc. NIPS*, 2017.

[45] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee. Learning to generate long-term future via hierarchical prediction. *arXiv preprint arXiv:1704.05831*, 2017.

[46] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proc. ICML*, pages 1096–1103. ACM, 2008.

[47] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *Proc. NIPS*, pages 613–621, 2016.

[48] G. Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

[49] W. F. Whitney, M. Chang, T. Kulkarni, and J. B. Tenenbaum. Understanding visual concepts with continuation learning. In *ICLR Workshop*, 2016.

[50] T. Xue, J. Wu, K. L. Bouman, and W. T. Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *Proc. NIPS*, 2016.

[51] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proc. CVPR*, 2011.

[52] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *Proc. ECCV*, 2014.

[53] Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee. Unsupervised discovery of object landmarks as structural representations. In *Proc. CVPR*, 2018.

[54] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *Proc. ECCV*, pages 94–108. Springer, 2014.

[55] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Learning Deep Representation for Face Alignment with Auxiliary Attributes. *PAMI*, 2016.

Appendix

We first present more detailed results on MAFL dataset comparing performance of different versions of our method. Then we show extended versions of figures presented in the paper. The sections are organized by the datasets used.

A MAFL

Training set →			CelebA			VoxCeleb	
K landmarks	Regression set	Thewlis [43] (warp)	warp	no warp	no warp (selfsup.)	warp	no warp
10	MAFL	7.95	6.29	4.89	4.69	—	—
20		—	4.70	3.54	—	4.97	3.65
30		7.15	4.24	3.23	3.08	5.98	4.17
50		6.67	4.31	3.33	3.90	4.98	3.59
10	CelebA	6.32	6.27	4.88	4.68	—	—
20		—	4.69	3.53	—	5.00	3.66
30		5.76	4.30	3.22	3.07	5.95	4.14
50		5.33	4.29	3.32	3.90	5.01	3.55

Table 2: **Results on MAFL face-landmarks test-set.** Varying number (K) of unsupervised landmarks are learnt on two training-sets — random-TPS warps on CelebA [24], and face-videos from the VoxCeleb [29]. These landmarks are regressed onto 5 manually-annotated landmarks in the MAFL [55] test set, using either CelebA or MAFL training sets, with and without random-TPS warps (for data-augmentation, as in [43]). Mean squared-error (MSE) normalised by the inter-ocular distance is reported.

B MAFL and AFLW Faces

Figure 8: Supervised linear regression of 5 keypoints (bottom rows) from 30 unsupervised (top rows) on MAFL (above) and AFLW (below) test sets. Centre of the white-dots correspond to the ground-truth location, while the dark ones are the predictions. The models were trained on random-TPS warped image-pairs; self-supervised peceptual-loss network was used.

C VoxCeleb

Figure 9: Training with video frames from VoxCeleb. **[rows top-bottom]:** (1) source image $\mathbf{x}$, (2) target image $\mathbf{x}'$, (3) generated target image $\Psi(\mathbf{x}, \Phi(\mathbf{x}'))$, (4) unsupervised landmarks $\Phi(\mathbf{x}')$ superimposed on the target image. The landmarks consistently track facial features.

D BBCPose

Figure 10: **Learning Human Pose.** 50 unsupervised keypoints are learnt. Annotations (empty circles) for 7 keypoints are provided, corresponding to — head, wrists, elbows and shoulders. Solid circles represent the predicted positions; Top rows show raw discovered keypoints which correspond maximally to each annotation; bottom rows show linearly regressed points from the discovered keypoints. **[above]:** randomly sampled frames for different actors **[below]:** frames from a video track.

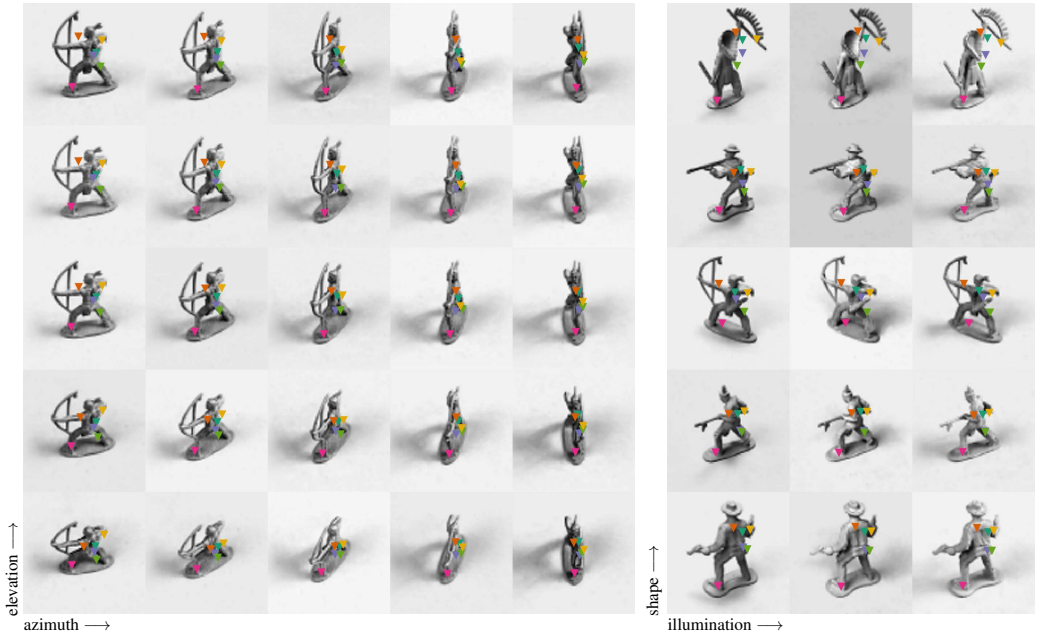
E Human3.6M

Figure 11: **Unsupervised Landmarks on Human3.6M.** Video of two actors (S1, S11) "posing", from the Human3.6M test set. **(rows)** (1) source, (2) target, (3) generated, (4) landmarks, (5) landmarks on frames from a different view, (6–7) landmarks on two views of the second actor. The landmarks consistently track the legs, arms, torso and head across frames, views and actors. However, the model confounds the frontal and dorsal sides.

F smallNORB 3D Objects: pose, shape, and illumination invariance

Object-category specific keypoint detectors are trained on the 5 categories in the smallNORB dataset — *human, car, animal, airplane,* and *truck*. Training is performed on pairs of images, which differ only in their viewpoints, but have the same object instance (or shape), and illumination.

Keypoints invaraint to viewpoint, illumniation, and object shape are visualised for object instances in the test set. The training set consists of only 5 object instances per category, yet the detectors generalise to novel object instances in the test set, and correspond to structurally similar regions across instances.




elevation ⟶
azimuth ⟶
shape ⟶
illumination ⟶
elevation ⟶
azimuth ⟶
shape ⟶
illumination ⟶
elevation ⟶
azimuth ⟶
shape ⟶
illumination ⟶

G Disentangling appearance and geometry

The generator substitutes the appearance of the target image ($\mathbf{x}'$) with that of the source image ($\mathbf{x}$). Instead of sampling image pairs ($\mathbf{x}, \mathbf{x}'$) with *consistent* style, as done during training, we sample pairs with *different* styles at inference, resulting in compelling transfer across different object categories — SVHN digits, Human3.6M humans, and AFLW faces.

Figure 12: **SVHN digits.** Target, source, and generated image triplets $\langle \mathbf{x}', \mathbf{x}, \Psi(\mathbf{x}, \Phi(\mathbf{x}')) \rangle$ from the SVHN test set. The digit shape is swapped out, while colours, shadows, and blur are retained.

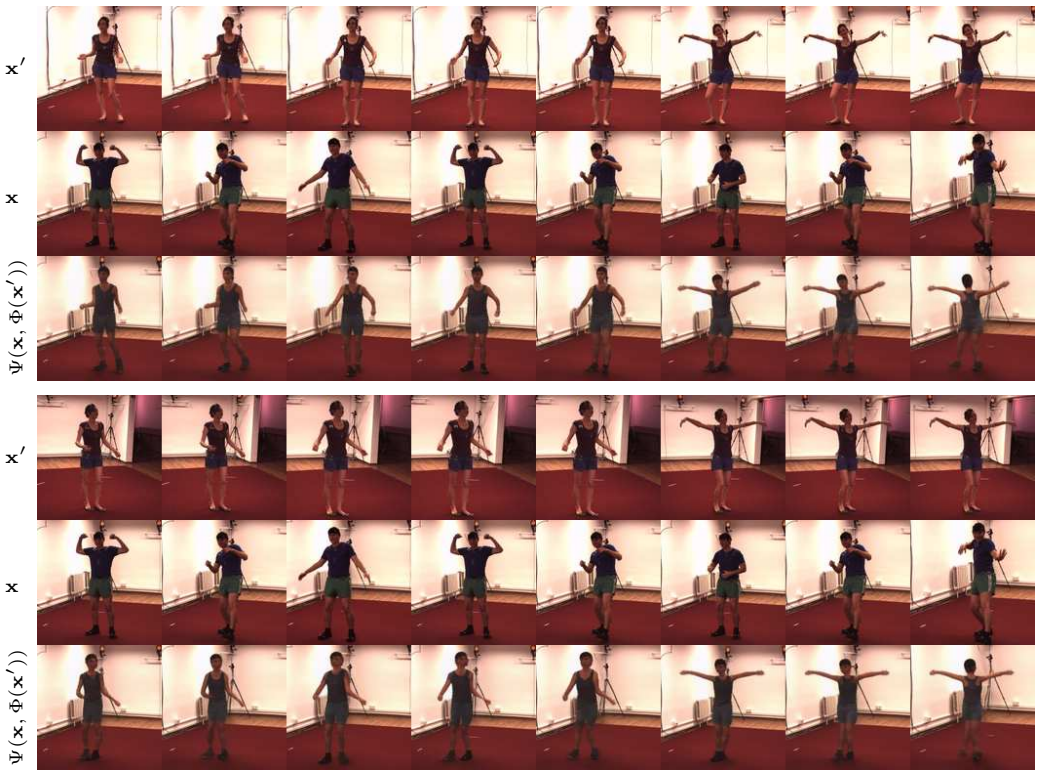


Figure 13: **Human3.6M humans.** Transfer across actors and viewpoints. **[top]:** different actors in various poses, imaged from the same viewpoint; the pose is swapped out, while appearance characteristics like shoes, clothing colour, and hat are retained. **[bottom]:** successful transfer even when the target is imaged from a different viewpoint (same poses as above).

Figure 14: **AFLW Faces.** The source image $\mathbf{x}$ is rendered with the pose from the target image $\mathbf{x}'$; the identity is retained.