Georgia State University

NOTICE OF ACADEMIC DISHONESTY

TO THE STUDENT: This is a notice that the below individual ("Initiator") believes that you have violated the University Policy on Academic Honesty and has recommended the penalty(ies) described below.

You will have 10 business days after this notice was emailed to you to submit a written appeal denying the charges and/or challenging the disciplinary penalty and providing any reason for the appeal. The appeal should be addressed to the college dean of the initiator listed below. If you do not respond within 10 business days, it will be assumed that you have accepted this finding of academic dishonesty. If you deny the charges, you will receive a notice of a college committee hearing to your student email address; the committee will make a finding of guilt or innocence.

If a disciplinary penalty has been recommended below, you have the right to challenge the disciplinary penalty (whether or not you deny the charges of academic dishonesty). The University Senate Committee on Student Discipline automatically reviews all recommendations for student disciplinary sanctions. Multiple findings of academic dishonesty may result in additional recommendations for disciplinary sanctions. Descriptions of penalties and a summary of review and processing procedures are contained on the reverse of this notice. This is a summary only; please refer to the Policy on Academic Honesty found in the Student Code of Conduct, available on the Dean of Students' website (codeofconduct.gsu.edu).

Student Motupalli, Sai Lohith **Panther ID** 002851852

Course Subject & Number CSC8260 **CRN #** 18415 **Term/Yr** Spring/2025 **Department** Computer science

Initiator Jingyu Liu **Department Chair** Armin Mikler **College** Arts and Science

Statement of finding of academic dishonesty by initiator (e.g., instructor):

For the final project of the course CSC 8260, Advanced digital image processing, this student submitted a project report which directly copied the results including tables and figures from a published paper, and used these results in his project presentation slides.

Academic penalty recommended: F for course

Disciplinary penalty recommended, if any: None

Initiator discussed Academic Dishonesty Charges with Student (circle one): **In Person** *or* **By Email** 5/7/2025 **Date**

Via Webex

Jingyu Liu Digitally signed by Jingyu Liu Date: 2025.05.09 15:06:23 -04'00' 5/7/2025

Initiator's Signature **Date**

Armin Robert Mikler Digitally signed by Armin Robert Mikler Date: 2025.05.09 16:27:46 -04'00'

Department Chair/Director Signature **Date**

5/13/2025

Date Received by College

Student Notified by College via Email:

5/15/2025 MB

Date Student Notified by College

ACADEMIC PENALTY: Academic penalties include assignment of a failing grade for the particular course requirement, or for the course itself, or for other tests or program assignments. They are set by the faculty member, in consultation with the department chair.

DISCIPLINARY PENALTY: Disciplinary penalties could include, but are not limited to, the following: suspension, expulsion, transcript annotation(s). Disciplinary penalties can be requested by the faculty member, in consultation with the chair.

SUMMARY OF REVIEW AND PROCESSING PROCEDURES: For the sake of brevity, the following summary is written from an "academic unit/college" perspective. Non-academic units (e.g., Testing Center) would substitute appropriate supervisory personnel at the respective levels. This is a summary only; please refer to the Policy on Academic Honesty found in the Student Code of Conduct, available on the Dean of Students' website (codeofconduct.gsu.edu).

1. The faculty member should discuss the incident with the student before filing a charge of academic dishonesty. The faculty member, in consultation with the department chair, prepares the Notice of Academic Dishonesty. The chair forwards the notice to the college dean, who sends the notification to the student's university email account.

2. The student must appeal in writing to the College Dean within 10 business days of the date the email was sent if the student wishes to deny the finding of academic dishonesty.

3. If the student does not appeal within 10 business days, the College Dean forwards the notice of academic dishonesty to the Dean of Students.

4. If the student appeals the charges, a College Hearing Committee conducts a hearing and reports its findings to the College Dean regarding guilt or innocence. If the student is found not guilty, the faculty member is notified to assign an appropriate grade. If the student is found guilty, the College Dean forwards the notice of academic dishonesty to the Dean of Students. The student may appeal a guilty finding to the Provost; the basis of the appeal must be as described in the Policy on Academic Honesty found in the Student Code of Conduct.

5. Any recommendation for a disciplinary penalty and a challenge of that disciplinary penalty submitted by the student, if any, is reviewed by the University Senate Committee on Student Discipline. Based on the committee's recommendation, the Provost makes a decision and takes action regarding any disciplinary sanction.

6. The Dean of Students maintains the disciplinary records on all findings of academic dishonesty and is responsible for forwarding notice of multiple findings to the Senate Committee on Student Discipline for review. Multiple findings may result in a disciplinary penalty even if one was not recommended by the faculty member.

Source: Senate Office, 9/7/95
Approved by the University Senate Committee on Admissions and Standards: 3/14/94
Approved by the University Senate: 11/3/94
Amended by the University Senate: 4/17/14
Revised for clarity by Office of Legal Affairs 8/25/2021

Support Document for Academic Dishonesty conducted by Sai Lohith Motupalli

For the final project, the project report and presentation slide submitted by the Student include **many parts copied from the published paper** 'Unsupervised Learning of Object Landmarks through Conditional Image Generation', by Tomas Jakab, Ankush Gupta, Hakan Bilen, Andrea Vedaldi, the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), https://proceedings.neurips.cc/paper_files/paper/2018/file/1f36c15d6a3d18d52e8d493bc8187cb9-Paper.pdf

The students' full report and slide are attached at the end of document, as well as the published paper. The following is examples of copied components.

In the Student report, the student uses the following figure and table as his own, which are **directly copy of the published paper**.

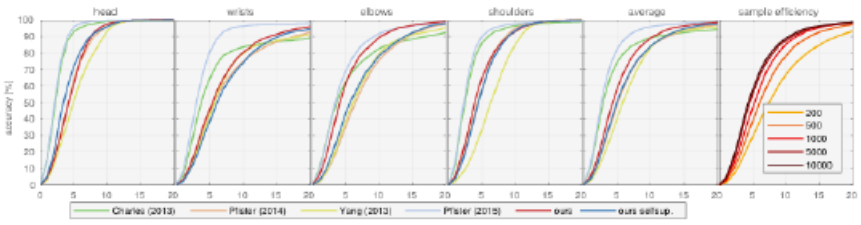


Fig. 1. The landmark detector ingests the image x to produce K landmark heatmaps y. It is composed of sequential blocks consisting of two convolutional layers each.

Method	K	MAFL	AFLW
Supervised			
RCPR [2]		–	11.60
CFAN [54]		15.84	10.94
Cascaded CNN [41]		9.73	8.97
TCDCN [57]		7.95	7.65
RAR [41]		–	7.23
MTCNN [56]		5.39	6.90
Unsupervised / self-supervised			
Thewlis [45]	30	7.15	–
	50	6.67	10.53
Thewlis [44](frames)	–	5.83	8.80
Shu † [38]	–	5.45	–
Zhang [55]	10	3.46	7.01
w/ equiv.	30	3.16	6.58
w/o equiv.	30	8.42	–
Wiles ‡ [51]	–	3.44	–
Ours, training set: CelebA			
loss-net: selfsup.	10	3.19	6.86
	30	2.58	**6.31**
	50	**2.54**	6.33
loss-net: sup.	10	3.32	6.99
	30	2.63	6.39
	50	2.59	6.35
Ours, training set: VoxCeleb			
loss-net: selfsup.	30	3.94	6.75
w/ bias	30	3.63	–
loss-net: sup.	30	4.01	7.10

Fig. 2. quantitative results

The student uses the following figure as his results in the project presentation, which are **directly copy of the published paper**.

Results - Human Poses & Objects

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.* [5]	65.9	47.9	66.5	76.8	64.1
Pfister *et al.* [34]	74.90	53.05	46.00	71.40	59.40
Yang *et al.* [53]	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

- Learns joint positions on video frames
- Tracks arms, head, shoulders in motion
- On SmallNORB: keypoints are invariant to lighting, shape, pose

Results - Faces

n supervised	Thewlis [45]	Ours selfsup
1	10.82	12.89 ± 3.21
5	9.25	8.16 ± 0.96
† 10	8.49	7.19 ± 0.45
100	—	4.29 ± 0.34
500	—	2.83 ± 0.06
1000	—	2.73 ± 0.03
5000	—	2.60 ± 0.00
All (19,000)	7.15	2.58± N/A

- Learns consistent keypoints (nose, eyes, mouth) without labels
- Beats some supervised models
- Regression to true landmarks: 2.58% MSE on MAFL (SOTA)

Unsupervised Learning of Object Landmarks through Conditional Image Generation

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

We propose a method for learning landmark detectors for visual objects (such as the eyes and the nose in a face) without any manual supervision. We cast this as the problem of generating images that combine the appearance of the object as seen in a first example image with the geometry of the object as seen in a second example image, where the two examples differ by a viewpoint change and/or an object deformation. In order to factorize appearance and geometry, we introduce a tight bottleneck in the geometry-extraction process that selects and distils geometry-related features. Compared to standard image generation problems, which often use generative adversarial networks, our generation task is conditioned on both appearance and geometry and thus is significantly less ambiguous, to the point that adopting a simple perceptual loss formulation is sufficient. We demonstrate that our approach can learn object landmarks from synthetic image deformations or videos, all without manual supervision, while outperforming state-of-the-art unsupervised landmark detectors. We further show that our method is applicable to a large variety of datasets — faces, people, 3D objects, and digits — without any modifications.

1 Introduction

There is a growing interest in developing machine learning methods that have little or no dependence on manual supervision. In this paper, we consider in particular the problem of learning, without external annotations, detectors for the landmarks of object categories, such as the nose, the eyes, and the mouth of a face, or the hands, shoulders, and head of a human body.

Our approach learns landmarks by looking at images of deformable objects that differ by acquisition time and/or viewpoint. Such pairs may be extracted from video sequences or can be generated by randomly perturbing still images. Videos have been used before for self-supervision, often in the context of future frame prediction, where the goal is to generate future video frames by observing one or more past frames. A key difficulty in such approaches is the high degree of ambiguity that exists in predicting the motion of objects from past observations. In order to eliminate this ambiguity, we propose instead to condition generation on two images, a source (past) image and a target (future) image. The goal of the learned model is to reproduce the target image, given the source and target images as input. Clearly, without further constraints, this task is trivial. Thus, we pass the target through a tight bottleneck meant to *distil the geometry of the object* (fig. 1). We do so by constraining the resulting representation to encode spatial locations, as may be obtained by an object landmark detector. The source image and the encoded target image are then passed to a generator network which reconstructs the target. Minimising the reconstruction error encourages the model to learn landmark-like representations because landmarks can be used to encode the *geometry* of the object,

[*] equal contribution.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

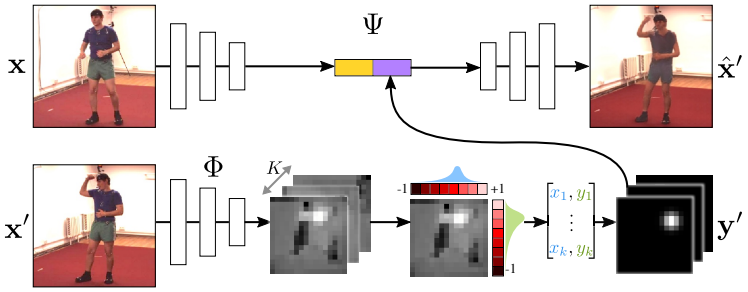


Figure 1: **Model Architecture.** Given a pair of source and target images $(\mathbf{x}, \mathbf{x}')$, the pose-regressor Φ extracts K heatmaps from $\mathbf{x}'$, which are then marginalized to estimate coordinates of keypoints, to limit the information flow. 2D Gaussians $(\mathbf{y}')$ are rendered 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.

which changes between source and target, while the appearance of the object, which is constant, can be obtained from the source image alone.

The key advantage of our method, compared to other works for unsupervised learning of landmarks, is the simplicity and generality of the formulation, which allows it to work well on data far more complex than previously used in unsupervised learning of object landmarks, *e.g*. landmarks for the highly-articulated human body. In particular, unlike methods such as [45, 44, 55], we show that our method can learn from synthetically-generated image deformations as well as raw videos as it *does not* require access to information about correspondences, optical-flow, or transformation between images.

Furthermore, while image generation has been used extensively in unsupervised learning, especially in the context of (variational) auto-encoders [22] and Generative Adversarial Networks (GANs [13]; see section 2), our approach has a key advantage over such methods. Namely, conditioning on both source and target images simplifies the generation task considerably, making it much easier to learn the generator network [18]. The ensuing simplification means that we can adopt the direct approach of minimizing a perceptual loss as in [10], without resorting to more complex techniques like GANs. Empirically, we show that this still results in excellent image generation results and that, more importantly, semantically consistent landmark detectors are learned without manual supervision (section 4). Project code and details are available at: `http://www.robots.ox.ac.uk/~vgg/research/unsupervised_landmarks/`

2 Related work

The recent approaches of [45, 44] learn to extract landmarks based on the principles of equivariance and distinctiveness. In contrast to our work, these methods are not generative. Further, they rely on known correspondences between images obtained either through optical flow or synthetic transformations, and hence, cannot leverage video data directly. 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 auto-encoders and restricted Boltzmann machines [14, 47, 15]. InfoGAN [6] 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*. [52] in variational auto-encoders, and by Whitney *et al*. [50] using a discrete gating mechanism to combine representations of successive video frames. Denton *et al*. [8] factor the pose and identity in videos through an adversarial loss on the pose embeddings. We instead design our bottleneck to explicitly shape the features to resemble the output of a landmark detector, without any adversarial training. Villegas *et al*. [46] also generate future frames by extracting a representation of appearance and human pose, but, differently from us, require ground-truth pose annotations. Our method essentially *inverts* their analogy network [36] to output landmarks given the source and target image pairs.

Several other generative methods [42, 40, 37, 48, 32] focus on video extrapolation. Srivastava *et al*. [40] employ Long Short Term Memory (LSTM) [16] networks to encode video sequences into fixed-length representation and decode it to reconstruct the input sequence. Vondrick *et al*. [48] propose a GAN for videos, also with a spatio-temporal convolutional architecture that disentangles foreground and background to generate realistic frames. Video Pixel Networks [20] 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 the spatial location of object landmarks.

A series of concurrent works propose similar methods for unsupervised learning of object structure. Shu *et al*. [38] learn to factor a single object-category-specific image into an appearance template in a canonical coordinate system, and a deformation field which warps the template to reconstruct the input, as in an auto-encoder. They encourage this factorisation by controlling the size of the embeddings. Similarly, Wiles *et al*. [51] learn a dense deformation field for faces but obtain the template from a second related image, as in our method. Suwajanakorn *et al*. [43] learn 3D-keypoints for objects from two images which differ by a known 3D transformation, by enforcing equivariance [45]. Finally, the method of Zhang *et al*. [55] shares several similarities with ours, in that they also use image generation with the goal of learning landmarks. However, their method is based on generating a single image from *itself* using landmark-transported features. This, we show is insufficient to learn geometry and requires, as they do, to also incorporate the principle of equivariance [45]. 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 $\mathbf{x}'$. We call $\mathbf{x}$ the source image and $\mathbf{x}'$ the target image and we use Ω to denote the image domain, namely the $H\times W$ lattice.

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 as a set of K object landmarks. As a first approximation, assume 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 Heatmaps bottleneck

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 via (spatial) $\mathrm{Softmax}$ 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 σ:

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

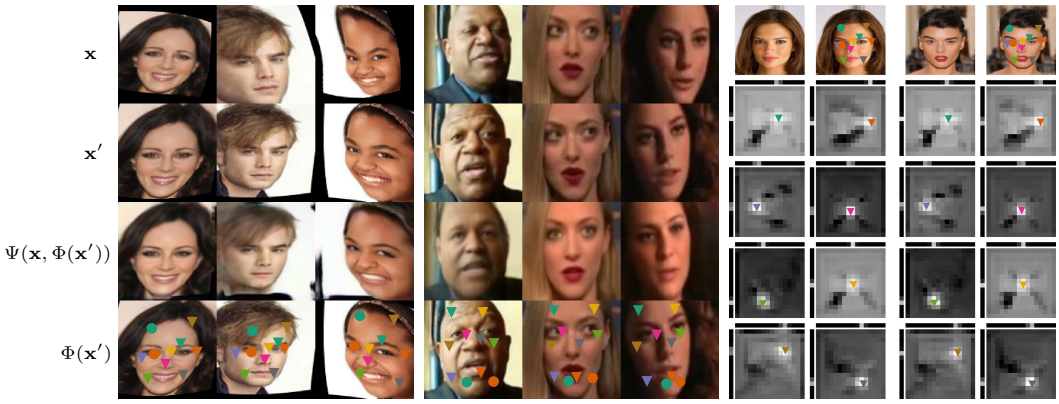


Figure 2: **Unsupervised Landmarks. [left]:** CelebA images showing the synthetically transformed source $\mathbf{x}$ and target $\mathbf{x}'$ images, the reconstructed target $\Psi(\mathbf{x}, \Phi(\mathbf{x}'))$, and the unsupervised landmarks $\Phi(\mathbf{x}')$. **[middle]:** The same for video frames from VoxCeleb. **[right]:** Two example images with selected (8 out of 10) 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 end result is a new tensor $\mathbf{y} = \Phi(\mathbf{x}) \in \mathbb{R}^{H \times W \times K}$ that encodes as Gaussian heatmaps the location of K maxima. Since it is possible to recover the landmark locations exactly from these heatmaps, this representation is 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 $\mathbf{x}'$ and generated $\Psi(\mathbf{x}, \Phi(\mathbf{x}'))$ images, as well as $\mathbf{x}'$ overlaid with the locations of the unsupervised landmarks $\Phi(\mathbf{x}')$. It also shows the heatmaps $S_u(\mathbf{x}; k)$ and marginalized separable softmax distributions on the top and left of each heatmap for $K = 10$ keypoints.

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 generative modelling, but rather to induce a representation $\mathbf{y}'$ of the object geometry for reconstructing a *specific* target image (as in an auto-encoder), a simpler method may suffice.

Inspired by the 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 [12, 1, 9, 19, 27, 30, 31]. 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 sub-network (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, namely colorising grayscale images [25].

n supervised	Thewlis [45]	Ours selfsup
1	10.82	12.89 ± 3.21
5	9.25	8.16 ± 0.96
† 10	8.49	7.19 ± 0.45
100	—	4.29 ± 0.34
500	—	2.83 ± 0.06
1000	—	2.73 ± 0.03
5000	—	2.60 ± 0.00
All (19,000)	7.15	$2.58 \pm$ N/A

Figure 3: **Sample Efficiency for Supervised Regression on MAFL. [left]:** Supervised linear regression of 5 keypoints (bottom-row) from 10 unsupervised (top-row) on MAFL test set. Centre of the white-dots correspond to the ground-truth location, while the dark ones are the predictions. Both unsupervised and supervised landmarks show a good degree of equivariance with respect to head rotation (columns 2, 4) and invariance to headwear or eyewear (columns 1, 3). **[right]:** MSE ($\pm\sigma$) (normalised by inter-ocular distance (in %)) 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. †: we outperform the previous state-of-the-art [45] with only 10 labelled examples.

We also test using a VGG-19 model pre-trained for image classification in ImageNet. All other networks are trained from scratch. 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

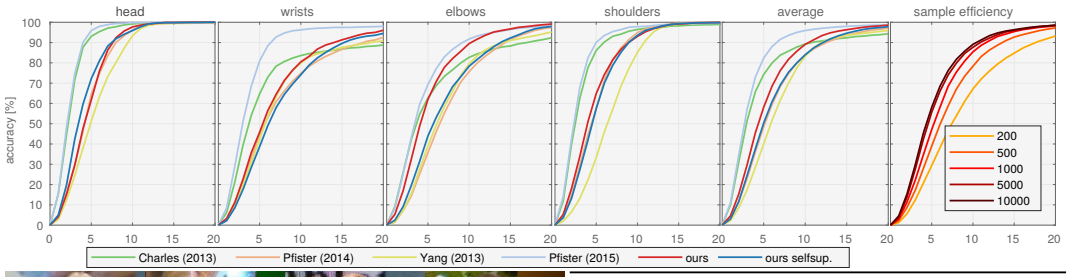
In section 4.1 we provide the details of the landmark detection and generator networks; a common architecture is used across all datasets. Next, we evaluate landmark detection accuracy on faces (section 4.2) and human-body (section 4.3). In section 4.4 we analyse the invariance of the learned landmarks to various nuisance factors, and finally in section 4.5 study the factorised representation of object style and geometry in the generator.

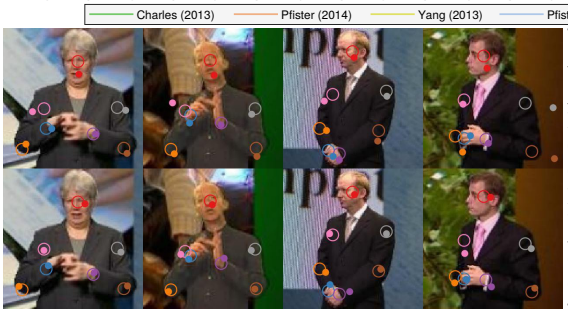
4.1 Model details

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 $16\times16\times K$ 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. A final convolutional layer regresses the three RGB channels with no non-linearity. All





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.* [5]	65.9	47.9	66.5	76.8	64.1
Pfister *et al.* [34]	74.90	53.05	46.00	71.40	59.40
Yang *et al.* [53]	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.

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 [21] with a weight decay of $5 \cdot 10^{-4}$. 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.

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 (TPS) [11, 49] warps g_1, g_2 to a given sample image $\mathbf{x}_0$. 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 [45] we exclude the images in MAFL [57] test-set from the training split and generate synthetically-deformed pairs as in [45, 55], but the transformations themselves are not required for training. We discount the reconstruction loss in the regions of the warped image which lie outside the original image to avoid modelling irrelevant boundary artefacts.

In the second case, $(\mathbf{x}, \mathbf{x}')$ are two frames sampled from a video. We consider VoxCeleb [28], a large dataset of face tracks, consisting of 1251 celebrities speaking over 100k English language utterances. We use the standard training split and remove any overlapping identities which appear in the test sets of MAFL and AFLW. Pairs of frames from the same video, but possibly belonging to different utterances are randomly sampled for training. By using video data for training our models we eliminate the need for engineering synthetic data.

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.

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

Quantitative results. We follow [45, 44] and use unsupervised keypoints learnt on CelebA and VoxCeleb to regress manually-annotated keypoints in the MAFL and AFLW [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 from the respective training sets. Model selection is done using 10% validation split of the training data.

We report results in terms of standard MSE normalised by the inter-ocular distance expressed as a percentage [57], and show a few regressed keypoints in fig. 3. Before evaluating on AFLW, we finetune our networks pre-trained on CelebA or VoxCeleb on the AFLW training set. We do not use any labels during finetuning.

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 $n = 10$ is already sufficient for results comparable to the previous state-of-the-art (SoA) method of Thewlis *et al*. [45], and that performance almost saturates at $n = 500$ (vs. 19,000 available training samples).

Vs. SoA. Table 1 compares our regression results to the SoA. We experiment regressing from $K{=}\{10, 30, 50\}$ unsupervised landmarks, using the self-supervised and the supervised perceptual loss networks; the number of samples n used for regression is maxed out ($= 19000$) to be consistent with previous works. On both MAFL and AFLW datasets, at 2.58% and 6.31% error respectively (for $K = 30$), we significantly outperform all the supervised and unsupervised methods. Notably, we perform better than the concurrent work of Zhang *et al*. [55] (MAFL: 3.16%; AFLW: 6.58%), while using a simpler method. When synthetic warps are removed from [55], so that the *equivariance constraint cannot be employed*, our method is significantly better (2.58% vs 8.42% on MAFL). We are also significantly better than many SoA *supervised* detectors [54, 41, 57] using only $n = 100$ supervised training examples, which shows that the approach is very effective at exploiting the unlabelled data. Finally, training with VoxCeleb video frames degrades the performance due to domain gap; including a bias in the linear regressor improves the performance.

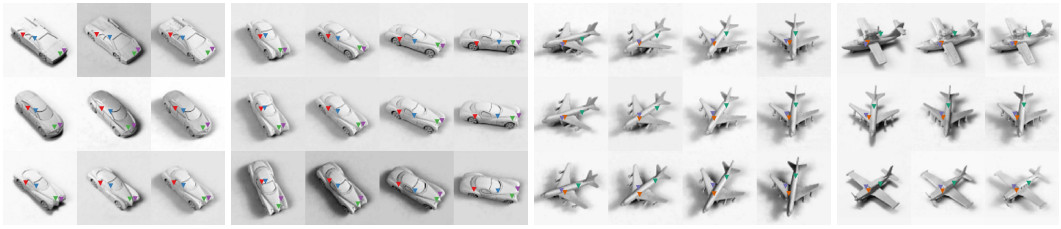
Method	K	MAFL	AFLW
Supervised			
RCPR [2]		–	11.60
CFAN [54]		15.84	10.94
Cascaded CNN [41]		9.73	8.97
TCDCN [57]		7.95	7.65
RAR [41]		–	7.23
MTCNN [56]		5.39	6.90
Unsupervised / self-supervised			
Thewlis [45]	30	7.15	–
	50	6.67	10.53
Thewlis [44](frames)	–	5.83	8.80
Shu † [38]	–	5.45	–
Zhang [55]	10	3.46	7.01
w/ equiv.	30	3.16	6.58
w/o equiv.	30	8.42	–
Wiles ‡ [51]	–	3.44	–
Ours, training set: CelebA			
loss-net: selfsup.	10	3.19	6.86
	30	2.58	**6.31**
	50	**2.54**	6.33
loss-net: sup.	10	3.32	6.99
	30	2.63	6.39
	50	2.59	6.35
Ours, training set: VoxCeleb			
loss-net: selfsup.	30	3.94	6.75
w/ bias	30	3.63	–
loss-net: sup.	30	4.01	7.10

Table 1: **Comparison with state-of-the-art on MAFL and AFLW.** K is the number of unsupervised landmarks. †: train a 2-layer MLP instead of a *linear* regressor. ‡: use the larger VoxCeleb2 [7] dataset for unsupervised training, and include a bias term in their regressor (through batch-normalization). Normalised %-MSE is reported (see fig. 3).

fc-layer (d) →	10	20	60	ours $K{=}30$
MAFL	20.60	21.94	28.96	**2.58**

loss →	ℓ_1	adv.+ ℓ_1	ℓ_2	adv.+ ℓ_2	content (ours)
MAFL ($K{=}30$)	3.64	3.62	2.84	2.80	**2.58**

Table 2: **Abalation Study. [left]:** The keypoint bottleneck when replaced with a low d-dimensional, $d = \{10, 20, 60\}$, *fully-connected* (fc) layer leads to significantly worse landmark detection performance (%-MSE) on the MAFL dataset. **[right]:** Replacing the *content* loss with ℓ_1, ℓ_2 losses on the images, optionally paired with an *adversarial* loss (*adv.*) also degrades the performance.



(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).

Ablation study. In table 2 we present two ablation studies, first on the keypoint bottleneck, and second where we compare against adversarial and other image-reconstruction losses. For both the settings, we take the best performing model configuration for facial landmark detection on the MAFL dataset.

Keypoint bottleneck. The keypoint bottleneck has two functions: (1) it provides a differentiable and distributed representation of the location of landmarks, and (2) it restricts the information from the target image to spatial locations only. When the bottleneck is replaced with a generic low dimensional fully-connected layer (as in a conventional auto-encoder) the performance degrades significantly This is because the continuous vector embedding is not encouraged to encode geometry explicitly.

Reconstruction loss. We replace our content/perceptual loss with ℓ_1 and ℓ_2 losses on generated pixels; the losses are also optionally paired with an *adversarial* term [13] to encourage verisimilitude as in [18]. All of these alternatives lead to worse landmark detection performance (table 2). While GANs are useful for aligning image distributions, in our setting we reconstruct a *specific* target image (similar to an auto-encoder). For this task, it is enough to use a simple content/perceptual loss.

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 [17]. 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 from multiple viewpoints. Image pairs are extracted by randomly sampling frames from 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 $K = 20$ and $K = 50$ keypoints are trained on Human3.6M and BBC-Pose respectively.

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 $\langle \mathbf{x}, \mathbf{x}', \Psi\left(\mathbf{x}, \Phi(\mathbf{x}')\right), \Phi(\mathbf{x}')\rangle$ 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 [33, 53] using 1k training samples (vs. 10k); furthermore, methods such as [35] are specialised for videos and

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.

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* (see supplementary material for visualisation of other object categories). 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 ($\pm 80^\circ$), 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 [29], AFLW 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 object landmark detectors. On faces, our method outperforms previous unsupervised as well as supervised methods for landmark detection. The method can also extend to much more challenging data, such as detecting landmarks of people, and diverse data, such as 3D objects and digits.

Acknowledgements. We are grateful for the support provided by EPSRC AIMS CDT, ERC 638009-IDIU, and the Clarendon Fund scholarship. We would like to thank James Thewlis for suggestions and support with code and data, and David Novotný and Triantafyllos Afouras for helpful advice.

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

[6] 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.

[7] J. S. Chung, A. Nagrani, and A. Zisserman. VoxCeleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

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

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

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

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

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

[13] 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.

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

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

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

[17] 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.

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

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

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

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

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

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

[29] 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.

[30] 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.

[31] 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.

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

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

[34] 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.

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

[36] S. E. Reed, Y. Zhang, Y. Zhang, and H. Lee. Deep visual analogy-making. In *Proc. NIPS*, 2015.

[37] 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.

[38] Z. Shu, M. Sahasrabudhe, A. Guler, D. Samaras, N. Paragios, and I. Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proc. ECCV*, 2018.

[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] S. Suwajanakorn, N. Snavely, J. Tompson, and M. Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. In *Proc. NIPS*, 2018.

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

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

[46] 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.

[47] 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.

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

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

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

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

[52] 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.

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

[54] 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.

[55] 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.

[56] 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.

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

Unsupervised Learning of Object Landmarks through Conditional Image Generation

Sai Lohith Motupalli
Master's in Computer Science
Georgia State University
smotupalli1@student.gsu.edu, Panther ID: 002851852

***Abstract*—In this project, I developed a method to detect key landmarks on objects like the eyes, nose, or joints—without using any labeled training data. Instead of manually annotating landmarks, I trained a model to learn them by observing pairs of images that show the same object in different poses or shapes. The model learns to separate an object's appearance from its geometry using a bottleneck mechanism that forces it to focus on spatial structure. Unlike other methods that rely on complex generative adversarial networks (GANs), my approach uses a simpler perceptual loss to guide training. This makes it easier to train and still achieves high-quality results. I tested the model on a wide range of datasets including faces, human bodies, 3D objects, and digits and it successfully learned meaningful landmarks in all cases, outperforming many existing unsupervised techniques.**

I. Introduction

A. Maintaining the Integrity of the Specifications

There's a growing need for machine learning methods that don't rely heavily on labeled data. In this project, I focused on detecting key landmarks on objects like eyes, nose, or shoulders without using any manual annotations. The idea is to train a model that learns from pairs of images showing the same object in different poses or angles. These pairs can be pulled from videos or created by slightly warping images.

Traditionally, some models try to predict future video frames from past ones, but that approach can be uncertain because motion is hard to predict. To avoid this, my method uses both a source image (for appearance) and a target image (for geometry) to train the model. The target image is passed through a bottleneck layer that forces the model to only focus on the object's structure, like where keypoints should be.

Then, the model combines the source image (appearance) and the compressed geometry from the target image to reconstruct the target. By minimizing the reconstruction error, the model naturally learns to place meaningful keypoints that represent the object's shape, while still using the source image for texture and style.

Detecting landmarks on visual objects such as eyes, nose, or body joints is a core task in many computer vision applications like face recognition, human pose estimation, and object tracking. Traditionally, achieving high accuracy in these tasks requires a large amount of manually annotated training data, which is both time consuming and expensive to obtain.

To address this limitation, there has been a growing interest in unsupervised learning techniques methods that can learn patterns and features without the need for manual labels. However, many existing unsupervised approaches are either too complex (e.g., using GANs or heavy constraints) or not general enough to handle diverse datasets like faces, bodies, or 3D objects.

In this study, I explored a simpler but powerful unsupervised method that learns landmarks by reconstructing a target image using the appearance from a source image and the geometry from the target image itself. By using a bottleneck layer to compress the target image into a set of heatmaps (representing keypoints), the model is forced to focus on learning object structure.

The key contribution of my work is showing that a single model, trained without labels or adversarial loss, can effectively learn landmarks across different object types and datasets, using just a perceptual loss for supervision.

B. ProjectRepo: https://github.com/Motupallisailohith/AIP_Project

II. Theory of related works

A. Theory Used

Previous methods for unsupervised landmark detection often rely on ideas like equivariance and distinctiveness meaning the model tries to find features that remain consistent under transformations. However, most of these approaches are not generative and depend on knowing how different images are related using things like optical flow or synthetic transformations which makes it hard to apply them directly to real video data. Our method avoids that dependency and can still benefit from equivariance, but doesn't require it as a core part of the model.

Traditionally, unsupervised representation learning has been done using autoencoders or restricted Boltzmann machines, and more recently, InfoGANs have been used to separate different factors (like pose and identity) in an image by forcing structure in the latent space. Our model also learns a structured representation, but instead of using an autoencoder or GAN, we use a conditional encoder-decoder architecture with a bottleneck that extracts keypoint-like features.

B. Related Works

Earlier studies have shown similar ideas — for example, Xue et al. used a variational autoencoder with a bottleneck, and Whitney et al. used a gating mechanism across video frames. Denton et al. separated pose and identity using GAN-based training. What makes our work different is that we don't use any adversarial training. Instead, we design our bottleneck to directly mimic how a landmark detector works, and we train using only a perceptual loss.

Other works like Villegas et al. also focus on future frame generation but need ground-truth pose labels. Our method goes one step further by learning directly from raw videos and generating landmarks without any labels.

There are also video-based generative models like Video Pixel Networks and LSTMs that try to model pixel-level dynamics across time. While powerful, they don't specifically focus on learning structure. In contrast, we focus on learning spatial keypoints as a compact representation of object structure.

Finally, some recent works come close to our approach. For example, Shu et al. learn to separate appearance and shape using a template-based decoder, and Wiles et al. learn dense deformation fields. Suwajanakorn et al. learn 3D keypoints using known 3D transformations. Zhang et al. also use generation for landmark learning, but they rely on transporting features within a single image, which we found isn't enough to capture geometry properly. What makes our model stand out is its simplicity and generality: it doesn't need optical flow or 3D info and works well with both synthetic image pairs and real video data.

III. MATERIALS AND METHODS

A. Data Explanation

The proposed method was tested on a diverse range of datasets: CelebA and MAFL: Contain face images with different poses and expressions. MAFL includes 5 annotated facial keypoints (used only for evaluation). BBCPose and Human3.6M: Feature full-body human poses, recorded as video sequences. BBCPose has 7 annotated points (e.g., head, wrists, elbows, shoulders). SmallNORB: Consists of 3D objects like cars and airplanes captured from multiple angles and lighting conditions. SVHN: Real-world digit images from house numbers. All images were resized to a resolution of 128x128. For video datasets, frame pairs were selected within a defined frame gap.

B. Preprocessing

Several preprocessing steps were carried out to ensure consistency across datasets: All images were resized to 128x128 pixels and normalized. For CelebA, synthetic image pairs were generated using Thin Plate Spline (TPS) warping to simulate different viewpoints or object deformations. For BBCPose and Human3.6M, frame pairs were sampled from the same video at different time intervals (typically 3–30 frames apart) to capture realistic motion. Regions lying outside the valid image boundary after warping were masked out to avoid penalizing the model on irrelevant data during training.

C. Image Analysis / Model Architecture

The model follows an encoder-bottleneck-decoder structure: Encoder (Keypoint Extractor): A convolutional neural network (CNN) processes the target image to produce K heatmaps, one for each landmark. Softmax is applied spatially to each heatmap, turning it into a probability map. The expected value (center of mass) of each heatmap gives a landmark coordinate. Each coordinate is then used to generate a 2D Gaussian blob centered on that location — forming the geometry bottleneck. Decoder (Image Generator): A separate CNN processes the source image to extract appearance features. These features are concatenated with the Gaussian keypoint blobs. A regressor network upsamples this combined representation to reconstruct the target image. Loss Function: The model is trained using a perceptual loss that compares the VGG-19 feature maps of the reconstructed and real target image. This allows the model to focus on structure rather than pixel-level differences.

D. Evaluation & Interpretation

Facial Landmark Detection Achieved 2.58% error on MAFL, outperforming supervised MTCNN and unsupervised Thewlis Keypoints are consistent across identity and expression Human Body Pose Learns meaningful joints (head, shoulders, wrists) Accuracy competitive with supervised models on BBCPose 3D Object Landmarking (SmallNORB) Keypoints are robust to pose, shape, and lighting Localizes semantically similar parts across object instances Disentangling Style & Geometry Successfully separates appearance from structure Can swap styles while preserving pose (e.g., pose of one digit + style of another)

I 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. I train category-specific detectors for K = 20 keypoints using image-pairs from neighbouring viewpoints and show results in fig. 6 for car and airplane (see supplementary material for visualisation of other object categories). 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 (80), 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.

IV. RESULTS

A. Facial Landmark Detection

Achieved 2.58% error on MAFL, outperforming supervised MTCNN and unsupervised Thewlis Keypoints are consistent across identity and expression

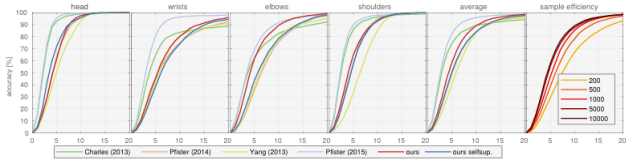


Fig. 1. The landmark detector ingests the image x to produce K landmark heatmaps y. It is composed of sequential blocks consisting of two convolutional layers each.

B. Human Body Pose

Learns meaningful joints (head, shoulders, wrists) Accuracy competitive with supervised models on BBCPose

C. 3D Object Landmarking (SmallNORB)

Keypoints are robust to pose, shape, and lighting Localizes semantically similar parts across object instances

D. Disentangling Style & Geometry

Successfully separates appearance from structure Can swap styles while preserving pose (e.g., pose of one digit + style of another)

V. Discussion and Conclusion

The results show that a simple bottleneck architecture combined with perceptual loss is enough to learn semantically rich landmarks without labels. The model generalizes across object types and domains without requiring specialized architecture or data. In this paper we have shown that a simple network trained for conditional image generation can be utilised to induce, without manual supervision, a object landmark detectors. On faces, our method outperforms previous unsupervised as well as supervised methods for landmark detection. The method can also extend to much more challenging data, such as detecting landmarks of people, and diverse data, such as 3D objects and digits.

A. Strengths:

Fully unsupervised Simple to train (no GANs, no flow) Works on faces, bodies, digits, 3D objects

B. Limitations:

Slight performance drop on challenging video frames Ambiguity in symmetrical structures (front vs. back)

C. Future Work:

Add temporal coherence Extend to 3D landmark localization Combine with tracking or segmentation tasks

Acknowledgment

I am very grateful for the support provided by EPSRCAIMSCDT, ERC638009 IDIU, and the Clarendon Fund scholarship. We would like to thank James Thewlis for suggestions and support with code and data, and David Novotný and Triantafyllos Afouras for helpful advice.

Method	K	MAFL	AFLW
Supervised			
RCPR [2]		–	11.60
CFAN [54]		15.84	10.94
Cascaded CNN [41]		9.73	8.97
TCDCN [57]		7.95	7.65
RAR [41]		–	7.23
MTCNN [56]		5.39	6.90
Unsupervised / self-supervised			
Thewlis [45]	30	7.15	–
	50	6.67	10.53
Thewlis [44](frames)	–	5.83	8.80
Shu † [38]	–	5.45	–
Zhang [55]	10	3.46	7.01
w/ equiv.	30	3.16	6.58
w/o equiv.	30	8.42	–
Wiles ‡ [51]	–	3.44	–
Ours, training set: CelebA			
loss-net: selfsup.	10	3.19	6.86
	30	2.58	**6.31**
	50	**2.54**	6.33
loss-net: sup.	10	3.32	6.99
	30	2.63	6.39
	50	2.59	6.35
Ours, training set: VoxCeleb			
loss-net: selfsup.	30	3.94	6.75
w/ bias	30	3.63	–
loss-net: sup.	30	4.01	7.10

Fig. 2. quantitative results

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

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

[6] J. S. Chung, A. Nagrani, and A. Zisserman. VoxCeleb2: Deep speaker recognition. In INTERSPEECH, 2018.

[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] J. Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In Constructive theory of functions of several variables. 1977.
[10] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In Proc. CVPR, 2016.
[11] G.E.Hinton and R.R.Salakhutdinov. Reducing the dimensionality of data with neural networks. Science, 313(5786):504–507, 2006. dependent pairwise relations. In Proc. NIPS, 2014.
[12] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8): 1735–1780, 1997.
[13] 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
[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In Proc. CVPR, 2017.
[15] J. Johnson, A. Alahi, and F. Li. Perceptual losses for real-time style transfer and super-resolution. In Proc. ECCV, 2016.

Unsupervised Learning of Object Landmarks through Conditional Image Generation

Presented by
Sai Lohith M
Panther ID: 002851852

Motivation

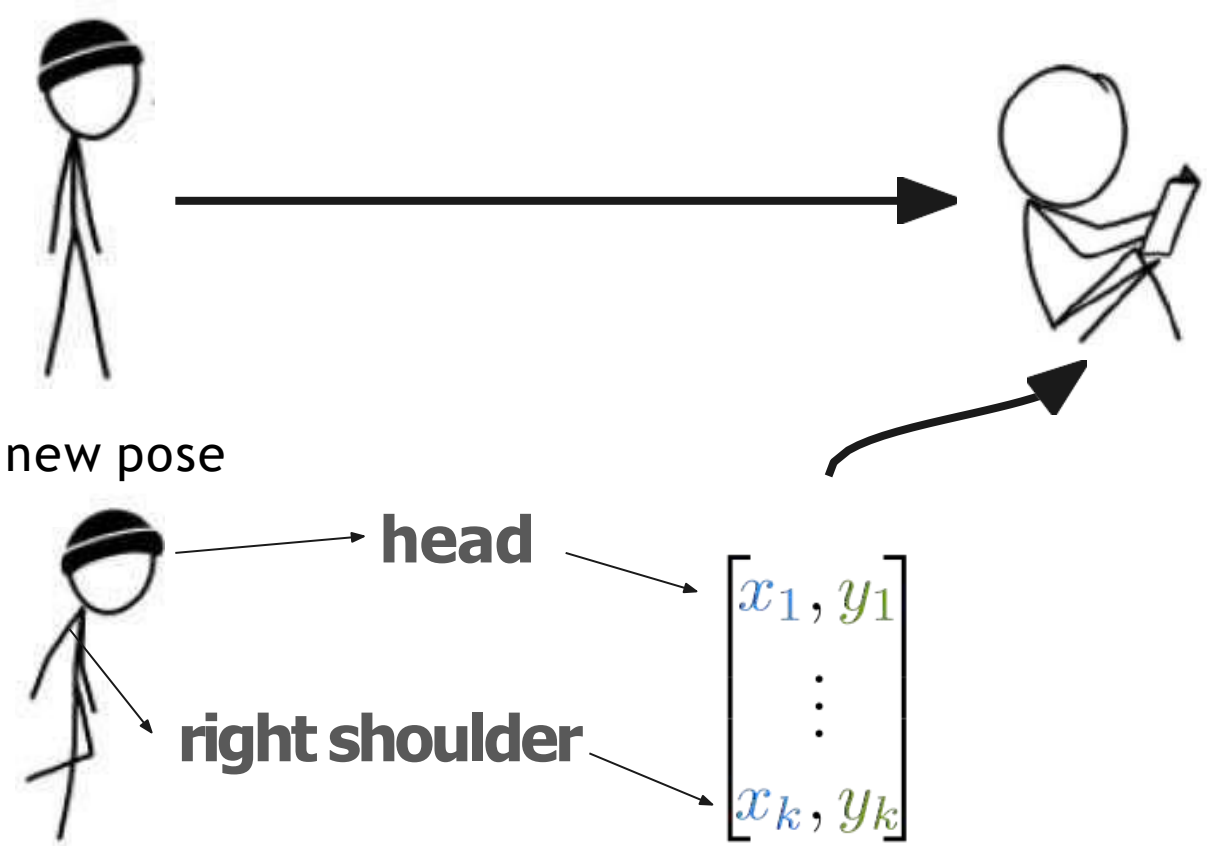
- Landmark detection (eyes, nose, joints) is crucial in vision tasks
- Manual annotation is expensive and not scalable
- **Goal**: Learn landmarks **without labels**
- Leverage videos or synthetic distortions

Intuition

Intuition

new pose

Intuition



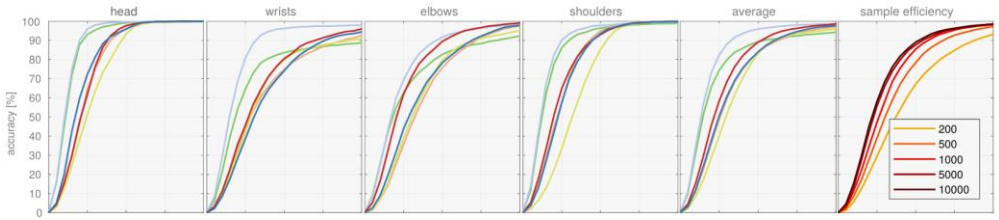
Model Architecture

Keypoint extractor (Phi):

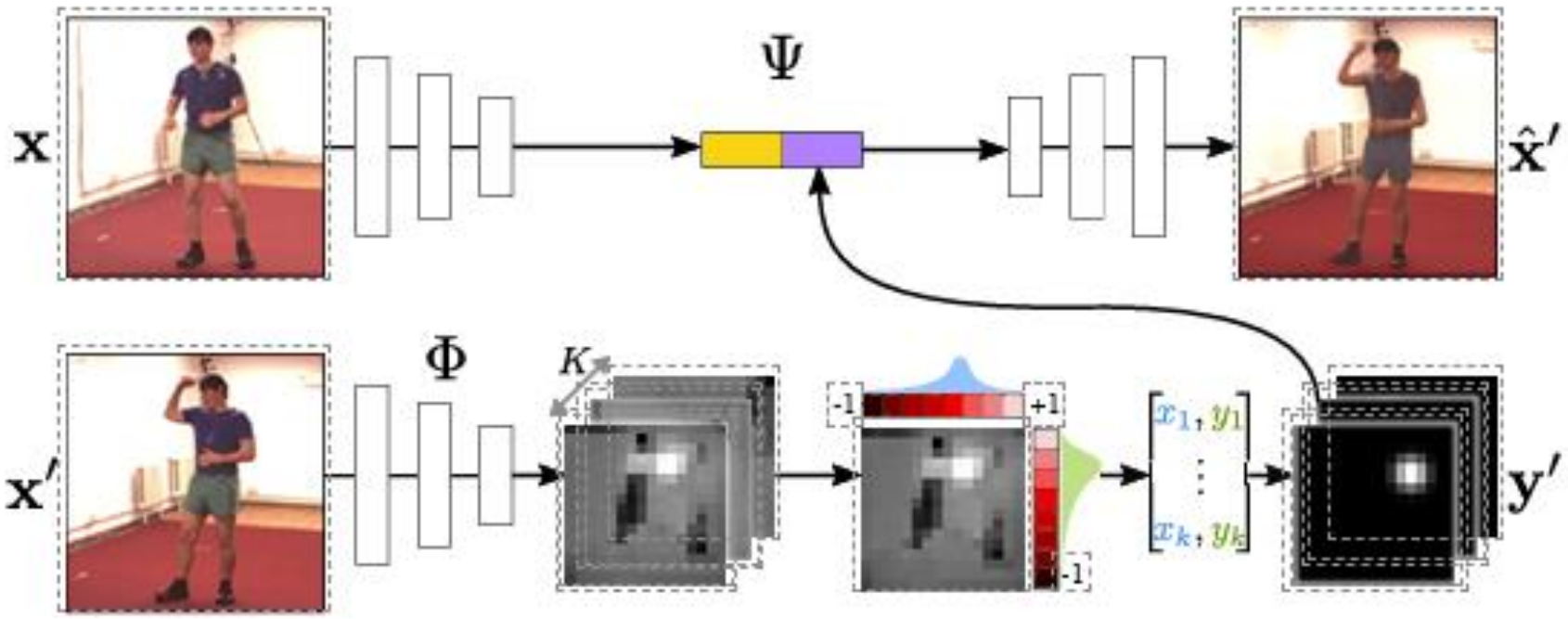
- Converts target image into K heatmaps
- Applies softmax to extract peak coordinates
- Converts to 2D Gaussian blobs

Image Generator (Psi):

- Combines source image + keypoint blobs
- Reconstructs the target image



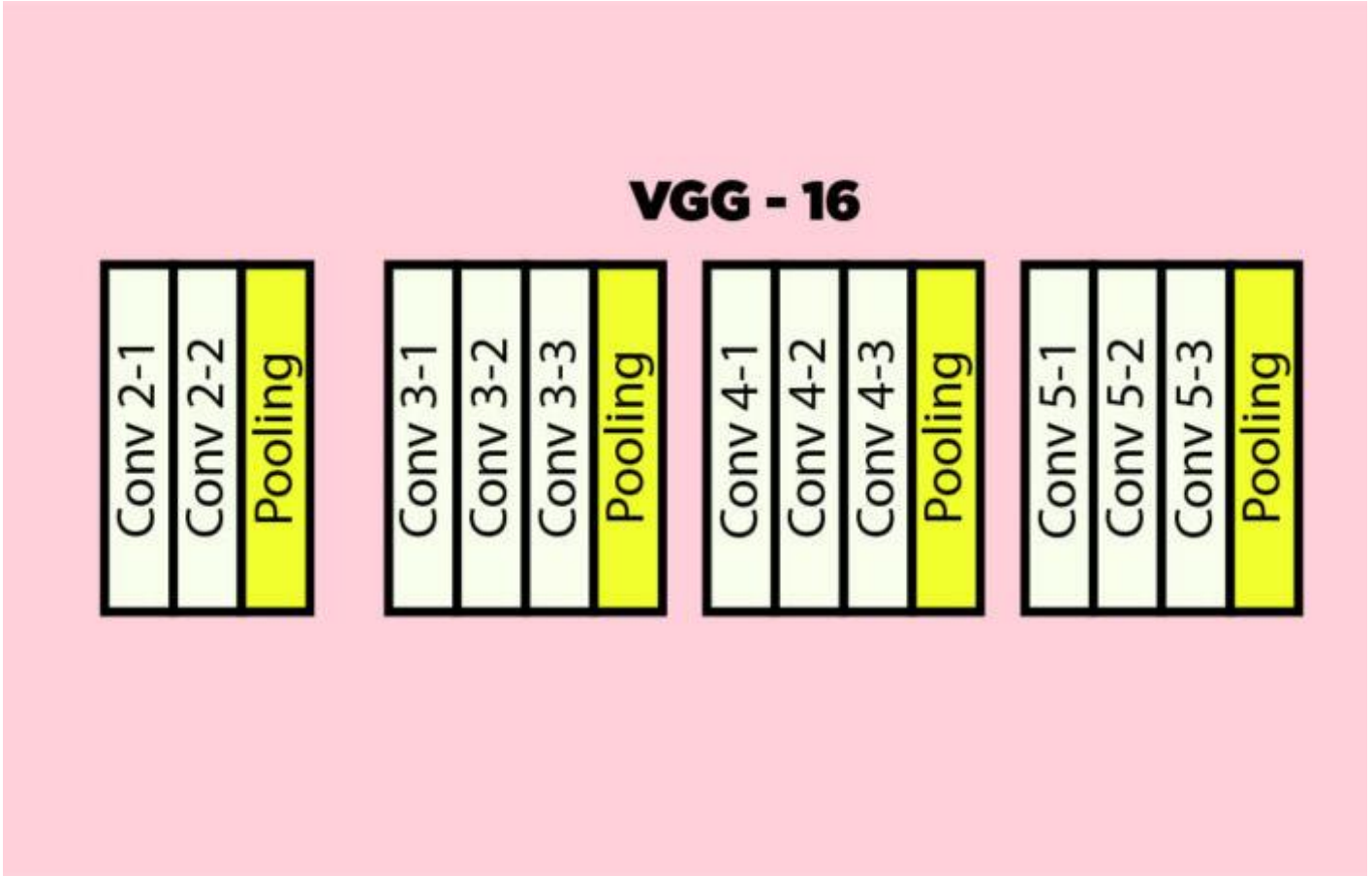
High-level-idea

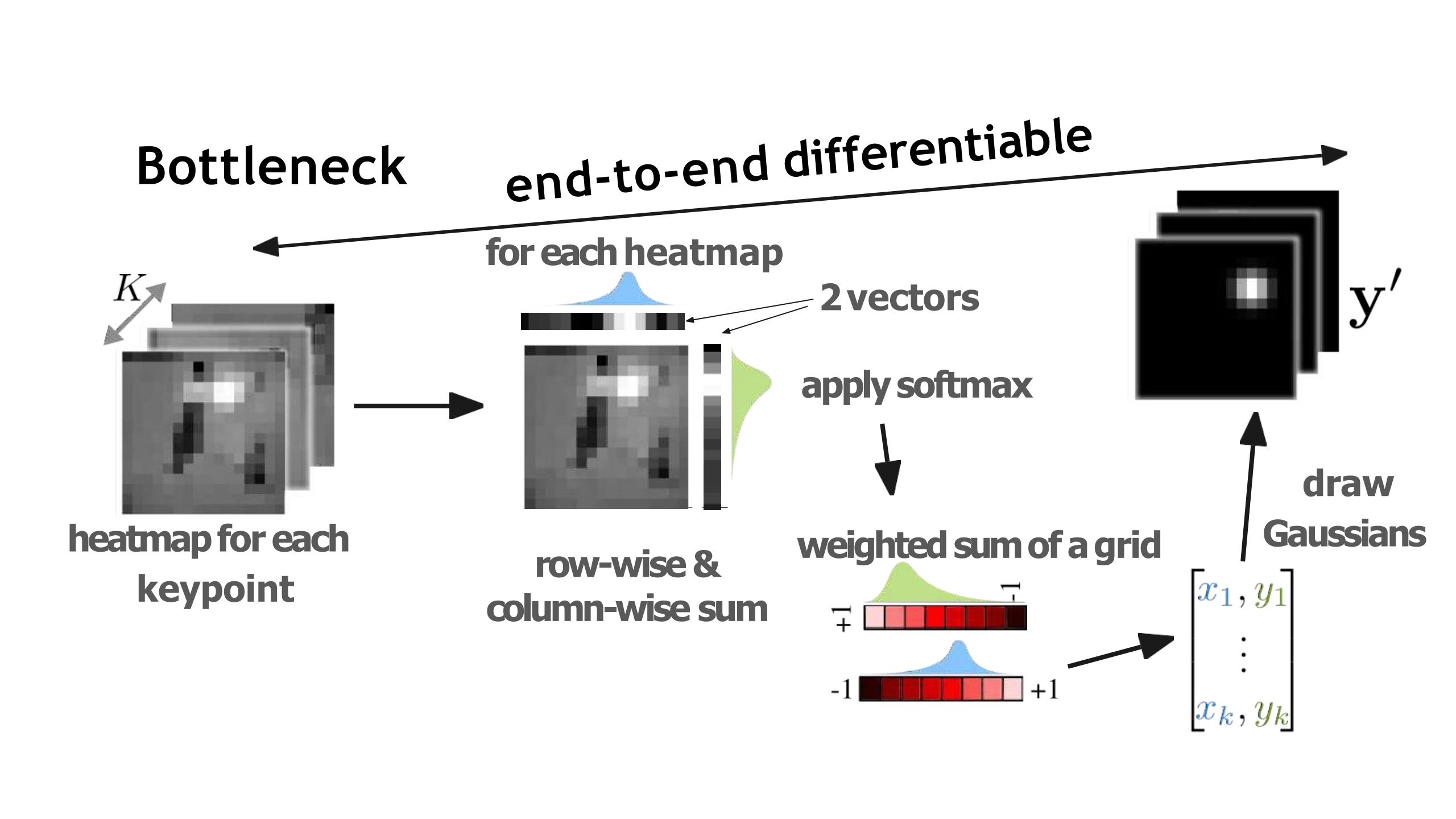


- Use image pairs: (source image, target image)
- Extract appearance from source, geometry from target
- Reconstruct target using both inputs
- Only way to succeed: learn meaningful landmarks!

Perceptual loss(using VGG features)

- Uses VGG network to compare images by feature maps
- Encourages structural similarity
- Avoids pixel-wise blurring
- Works better than L1/L2 loss or GANs for this task




Bottleneck
end-to-end differentiable
for each heatmap
2 vectors
apply softmax
K
heatmap for each keypoint
row-wise & column-wise sum
weighted sum of a grid
draw Gaussians
$\mathbf{y}'$
-1
+1
-1
+1
x_1, y_1
x_k, y_k

Results - Human Poses & Objects

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.* [5]	65.9	47.9	66.5	76.8	64.1
Pfister *et al.* [34]	74.90	53.05	46.00	71.40	59.40
Yang *et al.* [53]	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

- Learns joint positions on video frames
- Tracks arms, head, shoulders in motion
- On SmallNORB: keypoints are invariant to lighting, shape, pose

Results - Faces

n supervised	Thewlis [45]	Ours selfsup
1	10.82	12.89 ± 3.21
5	9.25	8.16 ± 0.96
† 10	8.49	7.19 ± 0.45
100	—	4.29 ± 0.34
500	—	2.83 ± 0.06
1000	—	2.73 ± 0.03
5000	—	2.60 ± 0.00
All (19,000)	7.15	$2.58 \pm$ N/A

- Learns consistent keypoints (nose, eyes, mouth) without labels
- Beats some supervised models
- Regression to true landmarks: 2.58% MSE on MAFL (SOTA)

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).

Some Inferences

Method	K	MAFL	AFLW
Supervised			
RCPR [2]		–	11.60
CFAN [54]		15.84	10.94
Cascaded CNN [41]		9.73	8.97
TCDCN [57]		7.95	7.65
RAR [41]		–	7.23
MTCNN [56]		5.39	6.90
Unsupervised / self-supervised			
Thewlis [45]	30	7.15	–
	50	6.67	10.53
Thewlis [44](frames)	–	5.83	8.80
Shu † [38]	–	5.45	–
Zhang [55]	10	3.46	7.01
w/ equiv.	30	3.16	6.58
w/o equiv.	30	8.42	–
Wiles ‡ [51]	–	3.44	–
Ours, training set: CelebA			
loss-net: selfsup.	10	3.19	6.86
	30	2.58	**6.31**
	50	**2.54**	6.33
loss-net: sup.	10	3.32	6.99
	30	2.63	6.39
	50	2.59	6.35
Ours, training set: VoxCeleb			
loss-net: selfsup.	30	3.94	6.75
w/ bias	30	3.63	–
loss-net: sup.	30	4.01	7.10

Disentangling Style and Geometry

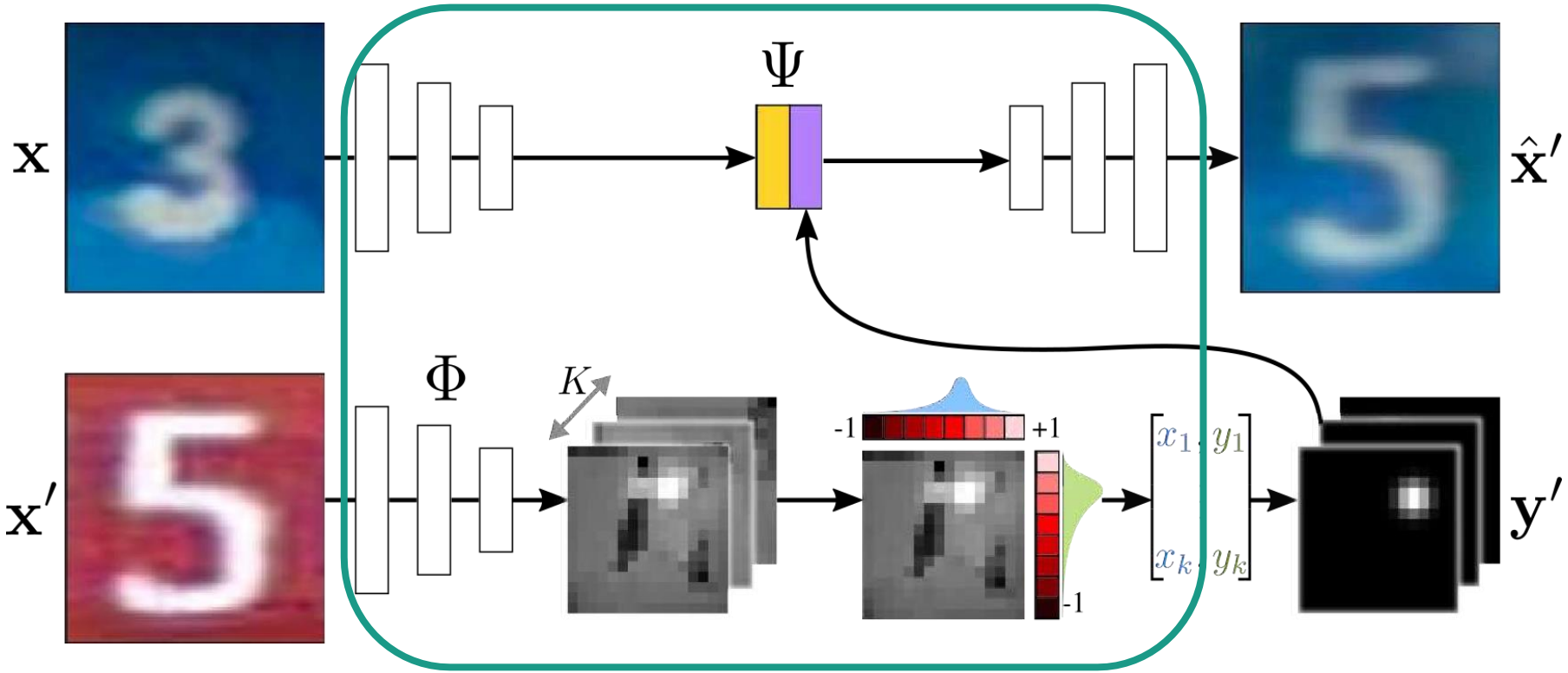


Image generation conditioned on spatial key points induces disentanglement of representations for style and geometry in the generator. Source image (x) imparts style (e.g. colour, texture), while the target image (x□) influences the geometry (e.g. shape, pose).

Street numbers

appearance

geometry

reconstruction

Human faces

appearance

geometry

reconstruction

Human pose

appearance

geometry

reconstruction

Strength & Limitations

- Simple pipeline, no labels or GANs
- Works across different domains
- Ambiguity in symmetric or frontal-back poses
- Slightly domain-sensitive

Future Work

- Add temporal smoothness or equivariance constraints
- Extend to 3D landmarks and real-time tracking
- Combine with action recognition or video segmentation

Thank you

Q & A

CSC 8260 Advanced Digital Image Processing
Spring 2025

Instructor:	Jingyu Liu
Course Time:	12:45 pm -14:30 pm Tuesday, Thursday Langdale Hall 429
Email:	iCollege email
Office Hours:	15:00-16:00pm, Tuesday, or by appointment Room 1818, 55 Park Pl
TA:	Omar Madjitov, Email: omadjitov1@student.gsu.edu

COURSE OBJECTIVES AND LEARNING OUTCOMES

This course will first review the fundamental concepts and processing of DIP from how digital images are created, stored, transformed, enhanced, to how they are used in daily life. And then we will introduce and study more advanced and recent techniques in DIP. Finlay you will apply these DIP techniques in one project with hand on experiences and write a project report and present your project.

In this course, you will learn about:	**By the end of the course, you will be able to:**
Theoretical foundation that supports each imaging processing step	Master the concepts of sampling rate, quantization, convolution, Fourier transformation, histogram. etc.
Review basic steps in DIP: image acquisition, image transformation, image enhancement, image compress, Image restoration, image segmentation, detection.	Master (describe, implement and evaluate) techniques of histogram equalization, intensity transformation, geometric transformation, spatial filtering, frequency filtering,
New techniques recently developed for image processing based on deep leaning, graph theory, transformer, for enhancement, segmentation, detection, etc.	Apply CNN, graph cut. GNN, transfer learning, vision Transformer, modeling techniques into relevant imaging process objectives.
Hands on experiences	Implement the techniques learned into one project
Presentation	**Deliver a written report and oral presentation**

TEXTBOOK

Rafael C. Gonzalez, Richard E. Woods, **Digital Image Processing, 4th edition**,

Other resources:

Lectures – Standford university **School of Engineering**

CS: Deep Learning for Computer Vision: http://cs231n.stanford.edu/index.html
Read relevant materials
CS: Machine Learning with Graphs https://snap.stanford.edu/class/cs224w-2023/index.html

Online lecture videos: Stanford University School of Engineering ;
https://www.youtube.com/playlist?list=PL3FW7Lu3i5JvHM8ljYj-zLfQRF3EO8sYv
Vision Transformer:
https://www.youtube.com/watch?v=vsqKGZT8Qn8

Relevant website materials
https://fpcv.cs.columbia.edu/
https://www.cl.cam.ac.uk/teaching/1718/CompVision/materials.html

Hand-on References:

1. Rafael C. Gonzalez, Richard E. Wood, Steven L. Eddins, Digital Image processing using MATLAB

2. Sandipan Dey: Hands-On Image Processing with Python: Expert techniques for advanced image analysis and effective interpretation of image data

LECTURE FORMAT: Lectures will be given in classes and examples of imaging processing will be given using MATLAB/python during the lectures. The related teaching materials will be uploaded into iCollege. Homework will be delivered into the iCollege Assignments. All assignments could be done using MATLAB or Python, and must be submitted into iCollege for grading. (no email submission is accepted)

REQUIREMENTS: Students are expected to **attend all classes**. The knowledge and skills you will gain on this course highly depend on your participation in class learning activities. Because of that, I expect you to attend all class sessions unless you are ill or have a valid reason for missing. I plan to track class attendance to help me understand how and when students are engaging in the course. If you are ill or have another valid reason for missing, please contact me by email in advance of absence.

Regular completion of all assignments and quizzes is absolutely essential to succeed in this course.

GRADING

Graded items include quizzes, homework, and project.

CSC8260	Quizzes	**30%**
	Homework	**30%**
	Final Project	**40%**

Final project needs to be presented in the class sessions, and the report needs to be written as IEEE conference format.

POLICIES

MAKE-UP POLICY

Make-up assignments will be offered in extremely limited circumstances. The instructor has the discretion to accept or refuse make-up requests. Explanations for missed examinations will need to be well documented.

POLICY ON LATE ASSIGNMENTS
All assignments must be uploaded to iCollege. Late quizzes will not be accepted. Other assignments will be considered late if they are posted after midnight on the due date unless an extension has been granted by the instructor. iCollege will accept assignments after the due date; however, assignments submitted more than 2 days late should be accompanied by documentation explaining the reason for lateness (e.g., a typed note of explanation from you, a scanned copy of a doctor's note, etc.). Late assignments will incur a 10% penalty per day.

POLICY ON ACADEMIC HONESTY
The University system policy on academic honesty is published in the *Faculty Affairs Handbook* and in *On Campus: The Undergraduate Co-Curricular Affairs Handbook* and is available to all members of the university community. The policy represents a core value of the university and all members of the university community are responsible for abiding by its tenets. Lack of knowledge of this policy is not an acceptable defense to any charge of academic dishonesty. All members of the academic community--students, faculty, and staff--are expected to report violations of these standards of academic conduct to the appropriate authorities. The minimal penalty for cheating in this class is the grade of F. The most common types of academic dishonesty are **plagiarism, cheating on assessments, and unauthorized collaboration.**

- **collaboration is allowed prior to preparation of actual material that is submitted for grade.**
- **Each student must work individually on his or her test. Any student found to be attempting to cheat on any test will receive a score of 0 for that test.**
- **Any attempt of getting or giving assistance in a test is considered cheating.**
- **It is the student's own responsibility to protect his or her work from being copied.**
- **No outside help is permitted.**
- **Plagiarized work is determined only by the instructor and is graded solely at the instructor's discretion.**

GSU Policy Prohibiting Students from Posting Instructor-Generated Materials on External Sites

The selling, sharing, publishing, presenting, or distributing of instructor-prepared course lecture notes, videos, audio recordings, or any other instructor-produced materials from any course for any commercial purpose is strictly prohibited unless explicit written permission is granted in advance by the course instructor. This includes posting any materials on websites such as Chegg, Course Hero, OneClass, Stuvia, StuDocu and other similar sites. Unauthorized sale or commercial distribution of such material is a violation of the instructor's intellectual property and the privacy rights of students attending the class, and is prohibited. This policy was approved by the GSU Faculty Senate on August 21, 2020.

COURSE EVALUATIONS
Your constructive assessment of this course plays an indispensable role in shaping education at Georgia State. Upon completing the course, please take time to fill out the online course evaluation.

ACCOMODATION FOR DISABILITY
Students who wish to request accommodation for a disability may do so by registering with the Office of Disability Services. Students may only be accommodated upon issuance by the Office of Disability Services of a signed Accommodation Plan and are responsible for providing a copy of that plan to instructors of all classes in which an accommodation is sought.

***** THE COURSE SYLLABUS PROVIDES A GENERAL PLAN FOR THE COURSE; DEVIATIONS MAY BE NECESSARY. *****

COURSE SCHEDULE

	Weekly Lecture Topics
Jan. 14, 16	Introduction to DIP and advanced DIP; Basic DIP operation (Sampling theory, Interpolation; Geometric transformation)
Jan. 21, 23	Intensity transformation for image enhancement (Homework 1)
Jan 28, 30	*Spatial filter, convolution,* and Frequency filter, (Quiz 1)
Feb. 4, 6	Frequency filter, 2D DFT, (Homework 2)
Feb. 11, 13	SIFT, (Quiz 2), Introduction of ANN
Feb. 18, 20	CNN , CNN architecture, (Homework 3)
Feb. 25, 27	CNN training, Image restoration, denoise (Homework 4)
Mar 4, 6	CNN application for super resolution, wavelet transformation (Quiz 3)

Mar. 11, 13	*image compression,* (Homework 5)
Mar. 17- 22	**Spring Break**
Mar. 25, 27	*Image segmentation, graph theory; graph cut,* GNN,
April 1, 3	CNN for image segmentation & detection*;* (Homework 6)
April. 8, 10	Transformer and Vision Transformer*; Morphological Image processing*
April. 15,17	Presentation
April 22, 24	Presentation
April 29	Report due