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

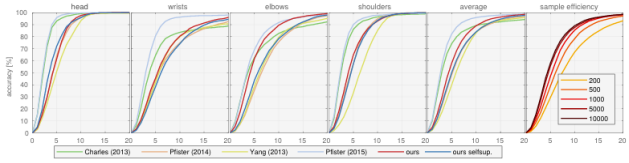

Fig. 1. The landmark detector ingests the image x to produce K landmark heatmaps y. It is composed of sequential blocks consisting of two convolutional layers each.

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.