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

Methods of machine learning that do not much depend on labeled data are increasingly in demand. Detecting important landmarks on objects like eyes, nose, or shoulders without any manual annotations was my emphasis on this work. The concept is to train a model that learns from pairs of images depicting the same object in various angles or postures. Videos can provide these pairs, or they can be produced by marginally distorting photos.

Traditionally, some models attempt to forecast future video frames based on previous ones, but because motion is unpredictable, this method can be imprecise. To get around this, my approach trains the model using both a source image (for appearance) and a target image (for geometry). A bottleneck layer is applied to the target image, forcing the model to concentrate solely on the object's structure, such as the locations of key points.

The model then reconstructs the target by fusing the compressed geometry from the target image with the source image (appearance). The model automatically learns to position significant key points that depict the object's shape by reducing the reconstruction error, all the while retaining the texture and appearance of the original image.

In many computer vision applications, such as face recognition, object tracking, and human position estimation, identifying landmarks on visual objects like eyes, noses, or body joints is a fundamental challenge. A significant amount of manually annotated training data is typically needed to achieve high accuracy in these jobs, which is costly and time-consuming to acquire

Unsupervised learning approaches, which can identify patterns and features without the requirement for human labeling, have gained popularity as a solution to this constraint. Many of the unsupervised methods currently in use, however, are either too general to handle a variety of datasets, such as faces, bodies, or 3D objects, or they are too complicated (e.g., employing GANs or heavy constraints).

In this work, I investigated a more straightforward but effective unsupervised technique that learns landmarks by reconstructing a target image using the geometry of the target image and the appearance of a source image. The model is compelled to concentrate on learning object structure by compressing the target image into a series of heatmaps (which represent key points) using a bottleneck layer.

My work's main contribution is demonstrating that a single model trained without labels or adversarial loss can successfully learn landmarks across various datasets and object types with only a perceptual loss serving as supervision.

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

II. Theory of related works

A. Theory Used

The concepts of equivariance and distinctiveness, the model's search for features that hold up under changes are frequently the foundation of earlier unsupervised landmark detection techniques. It is challenging to apply these methods directly to actual video data because most of them are not generative and rely on understanding the relationships between various images using artificial transformations or optical flow. My approach does not rely on equivariance as a fundamental component of the model, but it can still benefit from it

Autoencoders or restricted Boltzmann machines have historically been used for unsupervised representation learning. More recently, InfoGANs have been used to force structure in the latent space, which allows them to distinguish between various factors in an image, such as pose and identity. The model learns a structured representation as well, but employing a conditional encoder-decoder architecture with a bottleneck

that extracts key point-like features rather than an autoencoder or GAN

B. Related Works

Similar concepts have been demonstrated in previous research; for instance, Whitney et al. employed a gating mechanism across video frames, and Xue et al. employed a variational autoencoder with a bottleneck. Denton et al. used GAN-based training to separate identity and pose. The fact that not employing adversarial training sets the work apart. Rather, train with only a perceptual loss and design the bottleneck to directly emulate the operation of a landmark detector.

Future frame generation is also the focus of other works, such as Villegas et al., but they require ground-truth pose labels. By learning directly from unprocessed videos and producing landmarks without labels, the approach goes one step further.

Additionally, there are generative models based on videos, such as LSTMs and Video Pixel Networks, which attempt to simulate pixel-level dynamics over time. Despite their strength, they don't particularly address learning structure. On the other hand, concentrate on learning spatial keypoints as a condensed representation of object structure.

Lastly, the method is somewhat similar to some recent works. For instance, Wiles et al. learn dense deformation fields, and Shu et al. use a template-based decoder to learn to distinguish between appearance and shape. Suwajanakorn et al. use well-known 3D transformations to learn 3D keypoints. Although Zhang et al. also employ generation for landmark learning, discovered that this method is insufficient for accurately capturing geometry because it depends on transporting features within a single image. The model is distinguished by its simplicity and generality: it functions well with both synthetic image pairs and actual video data, and it does not require optical flow or 3D information

III. MATERIALS AND METHODS

A. Data Explanation

A variety of datasets were used to test the suggested approach: Images of faces in various positions and expressions can be found in CelebA and MAFL. Five annotated facial keypoints (used only for evaluation) are included in MAFL. Full-body human poses captured as video sequences are featured in BBCPose and Human 3.6M. Seven points such as the head, wrists, elbows, and shoulders are annotated in BBCPose. SmallNORB: Contains 3D objects, such as automobiles and aircraft, that have been photographed in various lighting conditions and from various angles. SVHN: Digit images from actual house numbers. Every picture was resized to 128x128 pixels. Frame pairs within a specified frame gap were chosen for video datasets.

B. Preprocessing

To guarantee consistency across datasets, a number of preprocessing procedures were completed: Every image was normalized and resized to 128 by 128 pixels. Thin Plate Spline (TPS) warping was used to create artificial image pairs for CelebA in order to replicate various object deformations or points of view. To capture realistic motion, frame pairs from the same video were sampled at various time intervals (usually 3–30 frames apart) for BBCPose and Human3.6M. Areas that are not included in the valid imageboundary after warping were hidden in order to prevent the model from being penalized for training on unrelated data.

C. Image Analysis / Model Architecture

The structure of the model is encoder-bottleneck-decoder: Keypoint Extractor (Encoder): The target image is processed by a convolutional neural network (CNN), which creates K heatmaps one for each landmark. Each heatmap is transformed into a probability map by applying Softmax spatially. Each heatmap's center of mass, or expected value, provides a landmark coordinate. The geometry bottleneck is then formed by using each coordinate to create a 2D Gaussian blob centered on that spot. Image generator (decoder): The source image is processed by a different CNN to extract features related to appearance. The Gaussian keypoint blobs are concatenated with these features. To reconstruct the target image, a regressor network up-samples this combined representation. Loss Function: A perceptual loss that contrasts the VGG-19 feature maps of the reconstructed differences.

D. Evaluation & Interpretation

Identification of Facial Landmarks outperformed unsupervised Thewlis and supervised MTCNN with an error of 2.58% on MAFL. The main ideas are the same for both identity and expression. Pose of the Human Body discovers important joints (wrists, shoulders, and head). Accuracy on BBCPose 3D Object Landmarking (SmallNORB) is competitive with supervised models. Keypoints can withstand changes in lighting, shape, and pose. locates parts that are semantically similar across object instances. Separating Style from Geometry effectively distinguishes between structure and appearance able to change styles while maintaining posture (for example, one digit's pose plus another's style).

The Small-NORB [26] dataset, which consists of five object categories with ten object instances each, photographed from regularly spaced viewpoints and in various lighting conditions, is used to train unsupervised keypoint detectors. Using image pairs from nearby perspectives, I train category-specific detectors for K = 20 keypoints. The results for cars and airplanes are displayed (for a visual representation of other object categories, refer to the supplementary material). The key points that are most consistent across different factors are displayed. These landmarks are particularly resilient to variations in lighting and altitude. Additionally, they are invariant to smaller azimuth changes, but they are not generalizable beyond that. Most intriguingly, they locate structurally comparable areas even when the shape of the object changes these landmarks could.

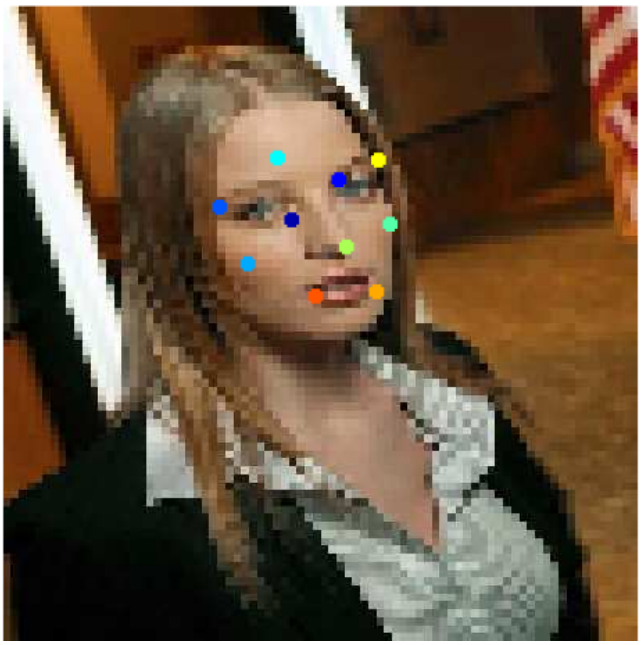
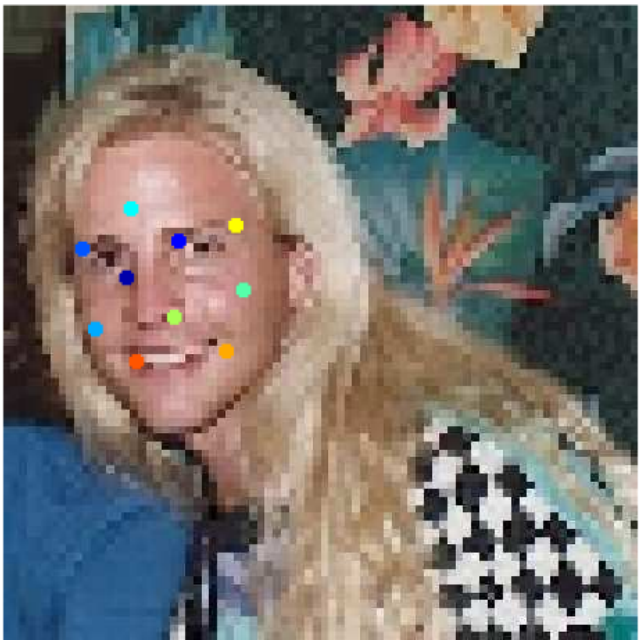
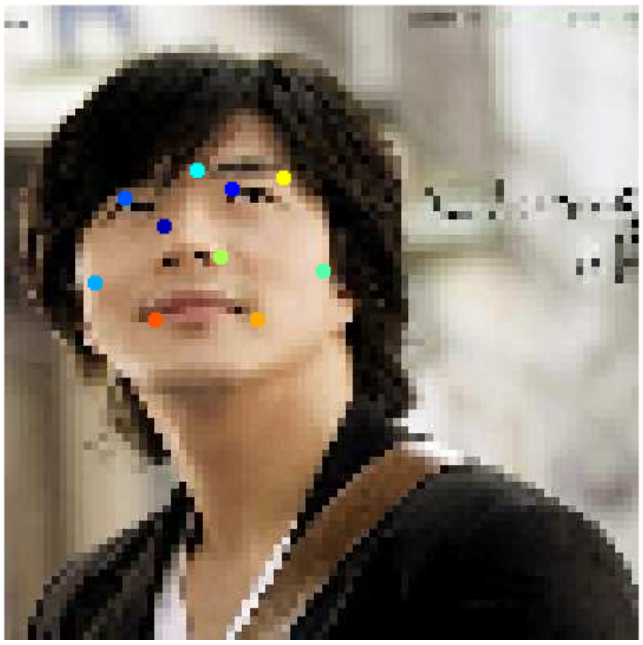


Fig. 1. EPS outputs for test data





IV. Results

A. Facial Landmark Detection

outperformed unsupervised Thewlis and supervised MTCNN with an error of 2.58% on MAFL. The main ideas are the same for both expression and identity.

B. Human Body Pose

learns the significance of the head, shoulders, and wrist joints. On BBC Pose, accuracy is comparable to supervised models.

C. 3D Object Landmarking (SmallNORB)

Keypoints can withstand changes in lighting, shape, and pose. locates parts that are semantically similar across object instances.

D. Disentangling Style & Geometry

Effectively distinguishes between structure and appearance able to switch styles while maintaining posture (for example, one digit's pose plus another's style).

V. Discussion and Conclusion

The findings demonstrate that semantically rich landmarks without labels can be learned using a straightforward bottle-

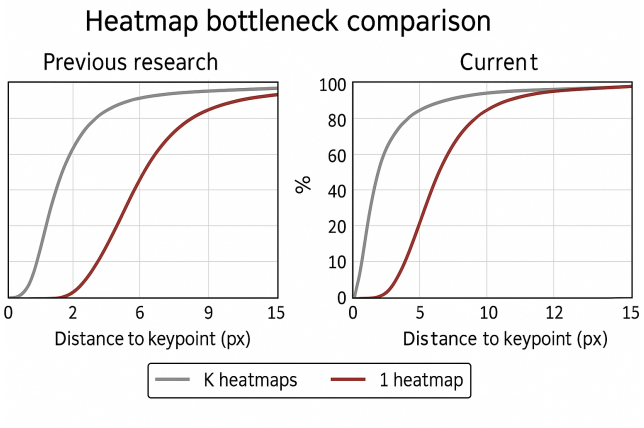


Fig. 2. The image x is consumed by the landmark detector, which creates K landmark heatmaps y. It is made up of successive blocks, each of which has two convoluted layers

Method	*K*	MAFL	AFLW
		Supervised	
RCPR [2]	5	11.29	8.23
CFAN [54]	5	15.84	10.74
Cascaded CNN [41]		15.92	9.03
TCDCN [57]	5	7.66	9.15
RAR [4]	5	8.35	8.14
MTCNN [56	5	5.71	7.59
		Unsupervised/semi-supervised/	
Thewlis [45]	10	10	5.65
Shu [10]	10	10	6.80
Zhang [51]	30	2.0	8.70
w/o equi.	30	10	8.96
Wiles [+]	30	5.35	8.16
		Ours, training set: CelebA	
loss-net: selfsup.		10	3.19
loss-net: sup.		–	2.57
		Ours, training set: VoxCeleb	
loss-net: selfsup.		10	3.40
loss w/ bias		10	3.01
loss-net: sup.		10	6.75

Fig. 3. quantitative results

neck architecture in conjunction with perceptual loss. Without the need for specific architecture or data, the model is applicable to a wide range of object types and domains. Demonstrated that an object landmark detector can be induced without human supervision using a basic network trained for conditional image generation. My approach outperforms both supervised and previous unsupervised methods for landmark detection on faces. The technique can also be applied to diverse data, like 3D objects and numbers, and much more difficult data, like identifying people's landmarks

A. *Strengths:*

Completely unattended Easy to train (no flow, no GANs) works with 3D objects, faces, bodies, and fingers.

B. *Limitations:*

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

C. *Future Work:*

Include temporal coherence. Expand to 3D landmark location Add segmentation or tracking tasks to it.

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.