Unsupervised Landmark Discovery Using Consistency Guided Bottleneck

Mamona Awan[1]
mamonaawan@yahoo.com
Muhammad Haris Khan[1]
muhammad.haris@mbzuai.ac.ae
Sanoojan Baliah[1]
sanoojan.baliah@mbzuai.ac.ae
Muhammad Ahmad Waseem[2]
msee20009@itu.edu.pk
Salman Khan[1,3]
salman.khan@mbzuai.ac.ae
Fahad Shahbaz Khan[1,4]
fahad.khan@mbzuai.ac.ae
Arif Mahmood[2]
arif.mahmood@itu.edu.pk

[1] Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE
[2] Information Technology University of the Punjab, Pakistan
[3] Australian National University, Australia
[4] Linköping University, Sweden

Abstract

We study a challenging problem of unsupervised discovery of object landmarks. Many recent methods rely on bottlenecks to generate 2D Gaussian heatmaps however, these are limited in generating informed heatmaps while training, presumably due to the lack of effective structural cues. Also, it is assumed that all predicted landmarks are semantically relevant despite having no ground truth supervision. In the current work, we introduce a consistency-guided bottleneck in an image reconstruction-based pipeline that leverages landmark consistency – a measure of compatibility score with the pseudo-ground truth – to generate adaptive heatmaps. We propose obtaining pseudo-supervision via forming landmark correspondence across images. The consistency then modulates the uncertainty of the discovered landmarks in the generation of adaptive heatmaps which rank consistent landmarks above their noisy counterparts, providing effective structural information for improved robustness. Evaluations on five diverse datasets including MAFL, AFLW, LS3D, Cats, and Shoes demonstrate excellent performance of the proposed approach compared to the existing state-of-the-art methods. Our code is publicly available at https://github.com/MamonaAwan/CGB_ULD.

1 Introduction

Object landmark detection is an important computer vision problem. It portrays important information about the shape and spatial configuration of key semantic parts in 3D space for deformable objects like human and animal faces [7, 10, 18, 23]. Many existing works have approached this problem in a fully-supervised manner [2, 5, 9, 13, 19, 29] which requires abundance of annotated images. Acquiring a large dataset of dense annotations for a particular object category may be infeasible. Therefore, the current work aims discovering object landmarks in an unsupervised way. Unsupervised learning of object landmarks is a challenging problem because the landmarks can express diverse configurations even for simple object categories like human faces. Also, recovering underlying mapping between spatial location and high-level semantic understanding of landmarks without involving human supervision is quite challenging. Finally, the consistency of landmark detection should not be compromised under viewpoint variations, and detected landmarks should capture the shape of the deformable object [23].

Existing approaches to unsupervised landmark detection either impose equivariance constraint to 2D image transformation [25, 26, 28], or leverage pre-text tasks such as (conditional) image generation [7, 23, 38]. For instance, [26] uses softargmax layer [34] to map the label heatmaps to a vector of points, and supervises the model with an equivariant error and a diversity constraint. Recently, Jakab *et al.* [7] proposed conditional image generation to guide learning of unsupervised landmark detection. They mapped the output of the softargmax layer to 2D Gaussian-like heatmaps using a *bottleneck* which is tasked with distillation of object geometry, and hence it learns structured embeddings. These heatmaps are then utilized to reconstruct the input image from its deformed version. The bottleneck is a crucial component in their pipeline as it guides the landmark detector to detect landmarks which are able to effectively reconstruct a deformed version of the same image. Using the same pipeline, Sanchez *et al.* [23] approached unsupervised landmark detection from a domain adaptation perspective via learning a projection matrix to adapt to new object categories. A problem inherent to these approaches is that they cannot alleviate the impact of noisy structural cues, which can affect robustness under pose variations (see Fig.1). We argue that a key reason is the naive formulation of the bottleneck. It assumes that, during training, all discovered landmarks by the detection network are equally meaningful under various variations. This is a strict assumption, as it is likely that at least some discovered landmarks will be noisy. The resulting noisy structural cues can potentially limit the reconstruction ability and affect the robustness of landmark detector, making it detect semantically irrelevant landmarks, lacking appropriate correspondence (see Fig.1).

In the current work, we address the aforementioned issues by introducing a *consistency-guided bottleneck* formulation that leverages landmark consistency to generate adaptive heatmaps. We rank the discovered landmarks based on their consistency and hence favour relatively consistent ones. We obtain pseudo-supervision via establishing landmarks correspondence across the images. It includes clustering landmarks after estimating their confidence in a KNN affinity graph. This consistency is then used to modulate the uncertainty of the landmark in the generation of adaptive heatmaps. As a result, the adaptive heatmaps favour consistent landmarks over their counterparts, thereby providing effective structural cues while reconstructing the input image. This, in turn, facilitates the landmark detector to produce semantically meaningful landmarks. [1] (see Fig.1).

Contributions: (1) We introduce a novel *consistency-guided bottleneck* formulation in the image reconstruction-based unsupervised landmark detection pipeline. It utilizes landmark consistency, a measure of affinity score with the pseudo-ground truth, for the generation

[1]Note that, the consistency-guided bottleneck facilitates detecting semantically meaningful landmarks and not semantic landmarks as such.

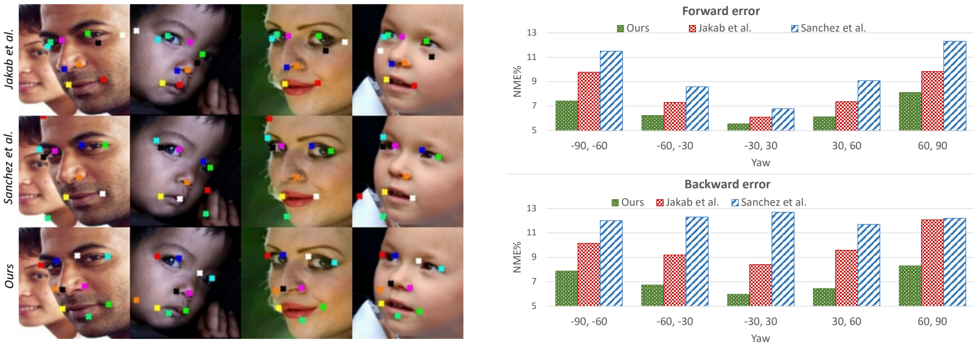

Figure 1: Left: Compared to ours, Jakab *et al.* [7] (top) and Sanchez *et al.* [23] (middle) are prone to discovering semantically irrelevant landmarks lacking appropriate correspondence across varying poses. Right: Comparison in terms of pose-wise NME(%) based on yaw-angles on the AFLW[12] dataset.

of adaptive heatmaps. Such heatmaps potentially encode better structural information to facilitate an improved discovery of semantically meaningful and stable points. **(2)** We propose a principled way of generating adaptive heatmaps in an unsupervised mode. We first rank landmarks based on their consistencies and then modulate their corresponding uncertainties in the 2D Gaussian heatmaps. **(3)** We also introduce pseudo-supervision via establishing landmark correspondence across images. **(4)** Comprehensive experiments and analysis are performed on five diverse datasets: MAFL, AFLW, LS3D, Cats, and Shoes. Our approach provides significant gains over the existing state-of-the-art methods.

2 Related Work

Unsupervised landmark detection methods can be broadly categorised into either imposing equivariance constraint to image transformations [26, 27, 28], or leveraging image reconstruction as a pre-text task [7, 8, 23]. In the absence of ground truth annotations, the equivariance constraint provides self-supervisory training signal. In particular, equivariance constraint requires representations across locations to be invariant to the geometric transformations of the image. Further constraints, based on locality [27, 28] and diversity [26] are introduced to avoid trivial solutions. The generative methods [1, 7, 8, 17, 23, 30, 31, 33] employ equivariance constraints rather implicitly by considering objects as a deformation of the shape template in-tandem with the appearance variation in a disentangled manner [4]. In [33], landmark discovery is formulated as an intermediate step of image representation learning. Similarly, [17] casts this as disentangling shape and appearance and introduced equivariance and invariance constraints into the generative framework. Wiles *et al.*[30] proposed a self-supervised framework to embed facial attributes from videos and then utilized those to predict landmarks. Most of these methods observe lack of robustness under pose variations.

Deep clustering methods employ clustering as pre-text task [3, 14, 15, 21, 32] to partition the images into different clusters and a classifier is trained to identify samples with same cluster id [14] or by using the cluster assignments as pseudo-labels [3, 21]. For unsupervised landmark discovery, Mallis *et al.* [18] recovers landmark correspondence via k-means clustering and utilized them to select pseudo-labels for self-training in the first stage. The pseudo-labels are

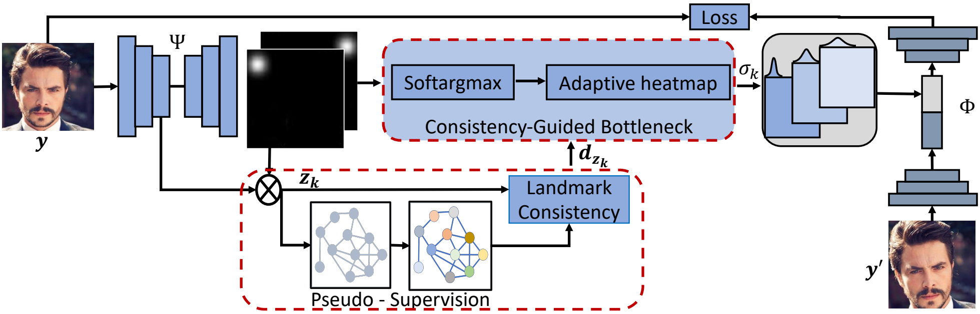

Figure 2: Overall architecture with consistency-guided bottleneck and pseudo-supervision.

used to learn a landmark detector in a supervised manner in the second stage. In contrast, we obtain pseudo-supervision to quantify landmark consistency. It is then used to modulate its 2D gaussian uncertainty in generating adaptive heatmaps. We do not use a dedicated feature head descriptor for learning landmark representations, and instead extract them directly from the encoder network. Moreover, we realize learning correspondence through clustering landmark representations after estimating their confidence in a KNN affinity graph.

3 Proposed Consistency Guided Bottleneck

We aim to train a model capable of detecting landmarks for an arbitrary object category, without requiring ground truth annotations. Similar to the prior works, we adopt an image generation based unsupervised landmark detection pipeline as shown in Fig. 2. It consists of a landmark detector network Ψ, and a generator network Φ. An important part of this pipeline is conditional image generation to guide the detection network in learning effective landmark representations. The object appearance in the first example image is combined with object landmark configuration in the second example image, where the two example images differ in viewpoint and/or object deformation. Heatmap bottleneck is a crucial component in this pipeline for factorizing appearance and pose. It has a softargmax layer and a heatmap generation process. Specifically, the network Ψ is terminated with a layer that ensures the output of Ψ is a set of k landmark detections. First, k heatmaps are formed, one for each landmark, then each heatmap is renormalised to a probability distribution via spatial Softmax and condensed to a point by computing the spatial expected value. Finally, each heatmap is replaced with a Gaussian-like function centred at landmark location with a particular standard deviation depending upon the consistency of that landmark. Although this unsupervised landmark detection pipeline shows encouraging results for some object categories, it struggles to detect semantically meaningful landmarks, especially under large pose variations (Figs. 1, & 4). We believe the key reason is the naive formulation of the bottleneck, comprising of a softargmax layer and a heatmap generation process. The bottleneck assumes that all predicted landmarks are equally meaningful (i.e. have same semantic relevance). It is likely that at least some of the landmark detections will be noisy, particularly in the absence of ground truth supervision. To address this, we introduce a *consistency-guided bottleneck* formulation that utilizes the landmark consistency towards generating adaptive heatmaps (Fig. 2).

3.1 Consistency of a Landmark

The consistency of a landmark is the proximity of its representation to an assigned pseudo-label which is a cluster centroid in our case. As such, it allows us to rank landmarks based on their consistency measures and hence favour relatively consistent ones over inconsistent ones. We obtain pseudo-supervision via establishing correspondence of landmarks across images. The process includes clustering the landmark representations after estimating their respective confidences in a KNN affinity graph. The consistency is then used to modulate the uncertainty of the landmark's 2D gaussian to generate adaptive heatmaps. Consequently, the adaptive heatmaps allow reducing the impact of noisy structural information (e.g., unstable landmarks) while reconstructing the image, which in turn allows the landmark detector to produce semantically meaningful and stable landmarks.

3.2 Obtaining Pseudo-Supervision

We obtain pseudo-supervision through establishing landmark correspondence across images. If two landmarks k^i and k^j in image i and image j correspond to the same semantic attribute (e.g. nose-tip), then their corresponding landmark representations $\mathbf{z}_k^i$, $\mathbf{z}_k^j$ should have the same pseudo-label. We realize this by clustering landmark representations after estimating their respective confidences in a KNN affinity graph. We use the landmark representations to construct a KNN affinity graph $G = (V, E)$. Where each landmark representation is a vertex belonging to V, and is connected to its $\mathcal{K}$ nearest neighbors, forming $\mathcal{K}$ edges belonging to E. The affinity between landmark k^i and landmark k^j is denoted as $s_{i,j}$, which is the cosine similarity between their representations $\mathbf{z}_k^i$ and $\mathbf{z}_k^j$.

Using this affinity graph, we intend to perform the clustering of landmark representations by estimating the confidence of each landmark representation. The confidence reflects whether a landmark representation (a vertex in the affinity graph) belongs to a specific semantic attribute. However, due to different variations in face appearance and pose, each landmark representation may have different confidence values even when they belong to the same semantic attribute (e.g., nose). For a landmark representation with high confidence, its neighboring landmark representations tend to belong to the same semantic attribute, while a landmark representation with low confidence is usually adjacent to the representations from the other landmarks. Based on this, it is possible to obtain the confidence $c_{\mathbf{z}_k^i}$ for each landmark representation vertex based on the neighboring labeled representations as [33],

$$c_{\mathbf{z}_k^i} = \frac{1}{|\mathcal{N}_{\mathbf{z}_k^i}|} \sum\nolimits_{\mathbf{z}_k^j \in \mathcal{N}_{\mathbf{z}_k^i}} (\mathbf{1}_{y^j = y^i} - \mathbf{1}_{y^j \neq y^i}).s_{i,j}, \tag{1}$$

where $\mathcal{N}_{\mathbf{z}_k^i}$ is the neighborhood of $\mathbf{z}_k^i$, y^i is the ground truth label of $\mathbf{z}_k^i$ and $s_{i,j}$ is the affinity between $\mathbf{z}_k^i$ and $\mathbf{z}_k^j$. However, due to training in unsupervised mode, we cannot use aforementioned expression to compute the confidence for a landmark representation, and instead use a pre-trained graph convolutional network [11] (GCN) to achieve the same.

With a pre-trained GCN, we can categorize the landmark representations based on their estimated confidences, to ultimately compute their cluster centroids. For a landmark representation vertex $\mathbf{z}_k^i$, neighbors with confidence larger than $\tilde{c}_{\mathbf{z}_k^i}$ show that they are more confident to belong to a certain cluster. Where $\tilde{c}_{\mathbf{z}_k^i}$ is the predicted confidence of $\mathbf{z}_k^i$. In this way, we assign each landmark representation to a cluster, and then compute the cluster-centroid by taking the mean of representations assigned to this cluster. We denote the number of cluster centroids by T and they are much larger than the number of landmarks K for capturing the intra-class variance in each semantic attribute [2]. So, each semantic attribute could occupy more than one cluster.

3.3 Quantifying landmark consistency

We quantify the consistency of a landmark by relating it to each of the cluster centroids. In particular, given a landmark feature representation $\mathbf{z}_k$, we compute its similarity with the representations of T cluster centroids and take the maximum similarity:

$$d_{\mathbf{z}_k} = \max_{t \in T} \langle \mathbf{z}_k, \mathbf{z}_t \rangle, \tag{2}$$

where $\langle .,. \rangle$ is the cosine similarity operator, $\mathbf{z}_t$ is feature representation of t^{th} cluster centroid, and $d_{\mathbf{z}_k}$ denotes the consistency of k^{th} landmark. We assume that, if a landmark representation $\mathbf{z}_k$ has higher similarity to its assigned cluster centroid, compared to another landmark representation, then it should be ranked higher in consistency compared to the other. We empirically observed that our model's learning strives to improve landmark consistencies. Landmark consistency is also related to the performance, so the improvement in landmark consistency is corroborated by the decrease in error.

3.4 Generating Adaptive Heatmaps

We propose to generate adaptive 2D Gaussian heatmaps, as opposed to fixed ones, as it is likely that at least some proportion of the discovered landmarks will be noisy. In fixed heatmaps, the uncertainties of 2D Gaussians have a same constant value. This is particularly suitable if all landmark positions are semantically relevant, lying very close to the true spatial location of the semantic attribute. It is only possible if those landmarks are either carefully annotated by a human or perhaps, produced by some state-of-the-art fully-supervised landmark detector. However, in unsupervised mode, this is rather unlikely and hence we propose to rank these landmarks via modulating their 2D Gaussian uncertainties, to alleviate the impact of noisy landmarks in heatmap generation process.

Let Ω denote the image grid of size $H \times W$. The landmark detector $\Psi(\mathbf{y})$ produces K heatmaps $S_u(\mathbf{y};k)$, $u \in \Omega$ one for each landmark $k = 1, ..., K$. Where u are the coordinates of a landmark. These heatmaps are generated as the channels of a $\mathbb{R}^{H \times W \times K}$ tensor. We re-normalize each heatmap to a probability distribution using spatial softmax [7]:

$$u_k^*(\mathbf{y}) = (\sum_{u \in \Omega} u e^{S_u(\mathbf{y};k)}) / (\sum_{u \in \Omega} e^{S_u(\mathbf{y};k)}). \tag{3}$$

In this work, we allow each 2D gaussian in a heatmap to reflect landmark's consistency. In particular, we modulate the uncertainty σ_k of 2D gaussian using the consistency $d_{\mathbf{z}_k}$ described in Eq. (2) as:$\sigma_k = 1/exp(d_{\mathbf{z}_k})$. Using this modulated uncertainty σ_k, we create *adaptive heatmaps* by forming a Gaussian-like function, centred at the location of discovered landmark k i.e. u_k.

$$\Psi_u(\mathbf{y};k) = \exp[-1/(2\sigma_k^2)||u - u_k^*(\mathbf{y})||^2] \tag{4}$$

This results in a new set of K adaptive heatmaps encoding the 2D Gaussian heatmaps the location of K maximas, however, with a modulated uncertainty of 2D Gaussians reflecting

[2]Note that, the value of T is determined by the KNN+GCN clustering itself, and is set to 80 in Kmeans clustering.

landmark consistency. As such, this alleviates the impact of noisy landmark detections, thereby highlighting the consistent ones. These adaptive heatmaps then become input along with the deformed image representation to the reconstructor network Φ. We observe that these adaptive heatmaps are a more informed encoding of spatial locations for the reconstructor network Φ. This in turn better facilitates the landmark detector Ψ in producing semantically meaningful landmarks across poses and object categories.

4 Experiments

Datasets: We validate our approach on human faces, cat faces and shoes. For human faces, we use CelebA [16] (comprising of more than 200k celebrity images), AFLW [12], and the challenging LS3D [2] (containing large poses). For CelebA, we exclude the subset of test images of MAFL [39], which are used to test our trained models. For AFLW, we used the official train and test partitions. For LS3D, we follow the same protocol as in [2, 18] and use 300W-LP [41] for training. For cat faces, we choose Cats Head dataset [37] (10k images). Following [23], we use 7,500 for training the landmark detector and the rest for testing. For Shoes, we choose UT-Zappos50k [35, 36] (50k images), and use train/test splits from [23].
Landmark detector network: We use the Hourglass architecture [20] as landmark detection network Ψ. To obtain landmark representation, we concatenate the feature maps from the last block of encoder (768-D) and then reduce their dimensions to 256 using 1x1 convolution. The network produces heatmaps of spatial resolution 32×32, which are converted into $K \times 2$ tensor with a softargmax layer. We use element-wise multiplication of 256-D feature maps and heatmaps, to get 256-D representations of landmarks. For a fair comparison and following [23], the landmark detector Ψ is initialised with the checkpoint, pre-trained on MPII. For details on image reconstruction network, we refer to the supplementary material.
Evaluation metrics: We use *forward* error [18, 23], *backward* error [23], and Normalised Mean-squared Error (NME), normalized by inter-ocular distance to report the performance.
Training details: We use $\mathcal{K} = 80$ in KNN affinity graph and use GCN to estimate confidences of the landmark representation vertices. In particular, we use a 1-layer pre-trained GCN on MS-Celeb-1M [6] dataset.We obtain pseudo-supervision after every 5 epochs. Our overall network architecture is trained for 145 epochs, with a learning rate of 1×10^{-4}, and a mini-batch size of 16 using Adam optimizer.

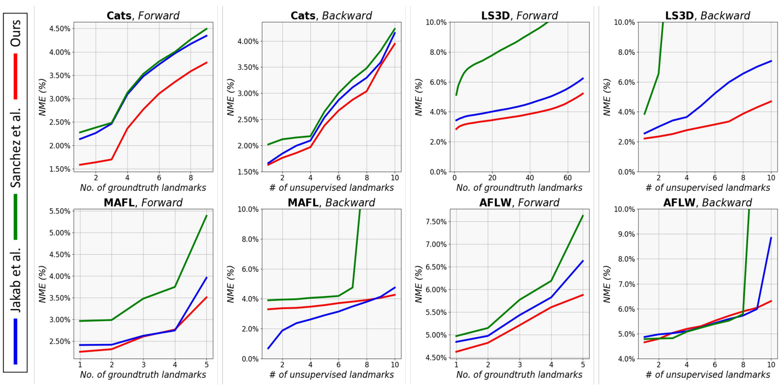

Figure 3: Cumulative error distribution (CED) curves for forward and backward errors.

Method		MAFL	AFLW
Sup	TCDCN [40]	7.95	7.65
	MTCNN [39]	5.39	6.90
Unsupervised	Thewlis [26](K=30)	7.15	-
	Jakab [7](K=10)†	3.32	6.99
	Jakab [7](K=10)‡	**3.19**	6.86
	Zhang [38]	3.46	7.01
	Shu [24]	5.45	-
	Shahasrabudhe [22]	6.01	-
	Sanchez [23]	3.99	6.69
	Mallis [18]	4.12	7.37
Ours	Baseline [7]	3.99	7.03
	Proposed	3.50	**5.91**

Table 1: Performance comparison with the SOTA on MAFL and AFLW in forward errors. †: uses the VGG-16 for perceptual loss, ‡: uses a pre-trained network for perceptual loss. Our method outperforms baseline by a notable margin in both datasets.

Method	Forw. Err.	Backw. Err.
Baseline[7]	5.38	7.06
Sanchez[23]	26.41	5.44
Mallis[18]	6.53	6.57
Ours	**5.21**	**4.69**

Method	Forw. Err.	Backw. Err.
Baseline[7]	4.53	4.06
Sanchez[23]	4.42	4.17
Ours	**3.76**	**3.94**

Table 2: Error comparison on (left) LS3D, (right) Cats Head datasets.

Comparison with the state-of-the-art (SOTA):
MAFL and AFLW: In the forward error evaluation (Tab. 1), our method outperforms the baseline by a notable margin in both MAFL and AFLW datasets. Furthermore, it provides a significant improvement over the recent top performing methods of [23] and [18] in both datasets. Our baseline is an in-house implementation of the existing pipeline. In backward error evaluation (Tab. 3), our approach demonstrates the best performance by achieving the lowest NME of 4.26% and 6.39% on MAFL and AFLW, respectively. See Fig. 3 for Cumulative Error Distribution (CED) curves. **LS3D, Cats and Shoes:** In LS3D, our method achieves the best performance in both forward and backward errors (Tab. 2 (left)), and detects semantically meaningful landmarks with improved correspondence (Fig. 4). On Cats Head, our method delivers improved performance compared to others in both forward and backward errors (Tab. 2 (right)), and despite variations (e.g., appearance and expressions) it discovers landmarks displaying improved correspondence across images (Fig. 4).
Stability Analysis: The stability of discovered landmarks is evaluated by measuring the error per landmark [23] as, $e_k = ||\Psi_k(A(\mathbf{y})) - A(\Psi_k(\mathbf{y}))||$, where A denotes a random similarity transformation. We report stability error, averaged over K=10 landmarks, in Tab. 4. Our method produces more stable landmarks than the competing approaches on most datasets.
Ablation Study and Analysis: See suppl. for a study on method specific hyperparameters.
On landmark consistency: We compare landmark consistencies via the consistency measure d during the training (Fig. 5). Our model learning strives to gradually improve landmark consistencies. In contrast, in baseline, the landmark consistencies remain almost the same during training. The landmark consistency also impacts (forward) error on test set and so in our case the improvement in landmark consistency is reflected by the decrease in the error. Fig. 6 (right) displays consistency-modulated heatmaps during training. Larger blob radius and higher redness indicate lower consistency.

Method	MAFL	AFLW
Baseline[7]	4.53	8.84
Sanchez[23]	14.74	25.85
Mallis[18]	8.23	-
Ours	**4.26**	**6.39**

Table 3: Backward errors comparison on MAFL and AFLW datasets.

Figure 4: Visual comparison of ours with Jakab et al. [7] and Sanchez et al. [23]. Our method discovers more semantically relevant landmarks and recovers improved correspondence.

Method	MAFL	AFLW	Cats Head	LS3D	Shoes
Baseline[7]	**2.16**	3.12	2.59	4.95	2.83
Sanchez[23]	8.78	7.56	2.58	21.3	2.45
Ours	2.37	**1.77**	**2.24**	**3.23**	**2.19**

Table 4: Stability errors for our method and the other two SOTA approaches.

Methods	AFLW				Cats Head			
	Epoch # 65		Epoch # 110		Epoch # 65		Epoch # 110	
	Silh.	CH	Silh.	CH	Silh.	CH	Silh.	CH
Kmeans	-0.042	38.85	-0.053	38.25	0.038	44.98	-0.04	43.36
KNN+GCN	**0.723**	**296.1**	**0.74**	**337.4**	**0.55**	**57.22**	**0.67**	**112.9**

Table 5: Quality of clustered landmark representations in our method using Silhouette coefficient and Calinski-Harabasz (CH) Index.

On landmark detector (Ψ) trained from scratch: Tab. 6 reports the performance of the baseline and our method when Ψ is trained from scratch instead of being initialized from a checkpoint. Our method outperforms baseline by notable margins.

Clustering Landmark Representations: Fig. 6 (left) visualizes the clustered landmark features using t-SNE. The features are well-separated into different classes, and hence facilitate effective correspondence establishment. We also observe clustering quality by KNN+GCN is much better than only Kmeans (see Tab. 5). **On pseudo-supervision:** Tab. 7 evaluates the strength of our novel *consistency-guided bottleneck formulation*, by replacing KNN affinity graph and refinement (KNN+GCN) with K-means for achieving pseudo-supervision.

Methods/Datasets	MAFL		AFLW		Cats Head	
	Fwd	Bwd	Fwd	Bwd	Fwd	Bwd
Baseline	6.27	16.6	9.02	26.3	14.1	44.4
Ours	3.92	8.49	6.85	11.7	4.1	3.41

Table 6: Performance of baseline and our method when the landmark detection network Ψ is trained from scratch.

Methods/Datasets	MAFL		Cats Head		LS3D	
	F	B	F	B	F	B
Baseline[7]	3.99	4.53	4.53	4.06	5.38	7.06
SOTA	3.99	4.53	4.42	4.06	5.38	6.57
Ours w/ KMeans	3.73	**3.90**	3.95	4.95	5.34	4.70
Ours w/ KNN+GCN	**3.50**	4.26	**3.76**	**3.94**	**5.21**	**4.69**

Table 7: Comparison when either using KNN+GCN or K-means for pseudo-supervision with baseline [7] and SOTA methods [7, 18, 23]. Red: Best, Blue: Second best.

We also plot the evolution of T during training for different pre-fixed values of $\mathcal{K}$ in KNN (Fig. 7). We see that, for a given $\mathcal{K}$, the value of T produced is less (approx by 20) than value of $\mathcal{K}$ throughout training.

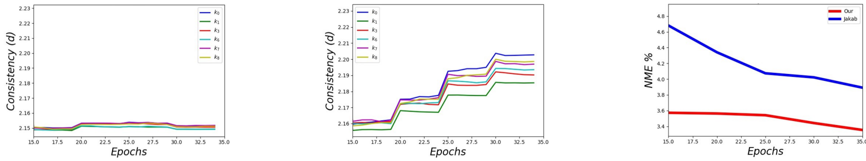

Figure 5: Comparison of average landmark consistency via d. (a) Baseline (Jakab et al.) (b) Ours (c) the impact of d on test forward error.

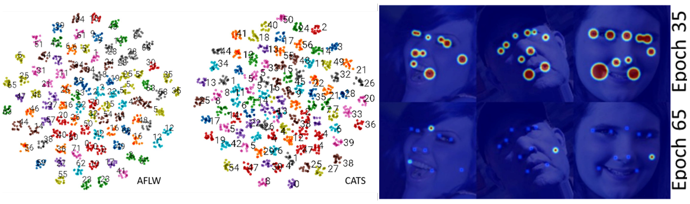

Figure 6: Left: Clustered features using tSNE with cluster ids. Right: Consistency-modulated heatmaps during training on AFLW. Larger blobs indicate lower consistency.

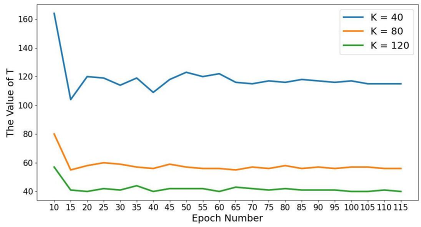

Figure 7: Evolution of T for three different pre-fixed $\mathcal{K}$ values.

Finally, we report both the forward and backward errors with different values of $\mathcal{K}$ (see Tab. 8). $\mathcal{K}$=80 used in our experiments, shows the best performance.

$\mathcal{K}$	40	80 (ours)	120
Forward (F) / Backward (B)	6.29/6.71	**5.91/6.39**	6.20/7.23

Table 8: Performance with different values of $\mathcal{K}$.

5 Conclusion

In this work, unsupervised landmark detection is improved by introducing a novel consistency-guided bottleneck. The landmark consistency is used for generating adaptive heatmaps. The consistency of a landmark is gauged by the proximity of its representation to the cluster center considered as pseudo label. Pseudo-supervision is established via landmark correspondence across multiple images. Extensive experiments on five publicly available datasets and a thorough analyses has demonstrated the effectiveness of the proposed approach. Excellent performance is observed compared to existing SOTA methods.

References

[1] Iaroslav Bespalov, Nazar Buzun, and Dmitry V Dylov. Brul\e: Barycenter-regularized unsupervised landmark extraction. *arXiv preprint arXiv:2006.11643*, 2020.

[2] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE ICCV*, pages 1021–1030, 2017.

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

[4] Zezhou Cheng, Jong-Chyi Su, and Subhransu Maji. On equivariant and invariant learning of object landmark representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9897–9906, 2021.

[5] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–388, 2018.

[6] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition, 2016.

[7] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. *NeurIPS*, pages 13520–13531, 2018.

[8] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8787–8797, 2020.

[9] Muhammad Haris Khan, John McDonagh, and Georgios Tzimiropoulos. Synergy between face alignment and tracking via discriminative global consensus optimization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3811–3819. IEEE, 2017.

[10] Muhammad Haris Khan, John McDonagh, Salman Khan, Muhammad Shahabuddin, Aditya Arora, Fahad Shahbaz Khan, Ling Shao, and Georgios Tzimiropoulos. Animalweb: A large-scale hierarchical dataset of annotated animal faces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6939–6948, 2020.

[11] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[12] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE ICCV workshops*, pages 2144–2151. IEEE, 2011.

[13] Abhinav Kumar, Tim K Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8236–8246, 2020.

[14] Dong Li, Wei-Chih Hung, Jia-Bin Huang, Shengjin Wang, Narendra Ahuja, and Ming-Hsuan Yang. Unsupervised visual representation learning by graph-based consistent constraints. In *European Conference on Computer Vision*, pages 678–694. Springer, 2016.

[15] Renjie Liao, Alexander Schwing, Richard S Zemel, and Raquel Urtasun. Learning deep parsimonious representations. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 5083–5091, 2016.

[16] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[17] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10955–10964, 2019.

[18] Dimitrios Mallis, Enrique Sanchez, Matthew Bell, and Georgios Tzimiropoulos. Unsupervised learning of object landmarks via self-training correspondence. *Advances in Neural Information Processing Systems*, 33, 2020.

[19] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. Direct shape regression networks for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5040–5049, 2018.

[20] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[21] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9359–9367, 2018.

[22] Mihir Sahasrabudhe, Zhixin Shu, Edward Bartrum, Riza Alp Guler, Dimitris Samaras, and Iasonas Kokkinos. Lifting autoencoders: Unsupervised learning of a fully-disentangled 3d morphable model using deep non-rigid structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[23] Enrique Sanchez and Georgios Tzimiropoulos. Object landmark discovery through unsupervised adaptation. *NeurIPS*, 32:13520–13531, 2019.

[24] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proceedings of the European conference on computer vision (ECCV)*, pages 650–665, 2018.

[25] Supasorn Suwajanakorn, Noah Snavely, Jonathan Tompson, and Mohammad Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. *arXiv preprint arXiv:1807.03146*, 2018.

[26] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *Proceedings of the IEEE international conference on computer vision*, pages 5916–5925, 2017.

[27] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. *arXiv preprint arXiv:1706.02932*, 2017.

[28] James Thewlis, Samuel Albanie, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6361–6371, 2019.

[29] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6971–6981, 2019.

[30] Olivia Wiles, A Koepke, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. *arXiv preprint arXiv:1808.06882*, 2018.

[31] Yinghao Xu, Ceyuan Yang, Ziwei Liu, Bo Dai, and Bolei Zhou. Unsupervised landmark learning from unpaired data. *arXiv preprint arXiv:2007.01053*, 2020.

[32] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5147–5156, 2016.

[33] Lei Yang, Dapeng Chen, Xiaohang Zhan, Rui Zhao, Chen Change Loy, and Dahua Lin. Learning to cluster faces via confidence and connectivity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[34] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European conference on computer vision*, pages 467–483. Springer, 2016.

[35] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE CVPR*, pages 192–199, 2014.

[36] Aron Yu and Kristen Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *Proc. of IEEE ICCV*, pages 5570–5579, 2017.

[37] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection-how to effectively exploit shape and texture features. In *ECCV*, pages 802–816. Springer, 2008.

[38] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2694–2703, 2018.

[39] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, pages 94–108. Springer, 2014.

[40] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):918–930, 2015.

[41] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3d solution. *CoRR*, abs/1511.07212, 2015. URL http://arxiv.org/abs/1511.07212.