Supplementary material for "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

Landmark detector pre-training. For a fair comparison and following [9], the landmark detector Ψ in our method, baseline [4], and others with similar pipeline [9] is initialised with the same checkpoint, pre-trained on MPII. Similarly, the VGG-16 network (in the reconstructor) is pre-trained on ImageNet for our approach, baseline [4] and [9].

Image reconstruction network. For image reconstruction, we adapt from architectures typically used for image-to-image translation [3], face synthesis [2, 7] and neural transfer [5]. We provide it with an image $\mathbf{y}'$ of resolution 128×128, where $\mathbf{y}'$ is the deformed version of original image $\mathbf{y}$. We create this deformed image $\mathbf{y}'$ by applying random similarity transformations over image $\mathbf{y}$. These transformations include scaling, rotation and translation. We then proceed by first applying two downsampling convolutions that bring the number of features to 256, and then concatenate the adaptive heatmaps with the downsampled image tensor to pass it through a set of 6 residual blocks. Finally, we apply two spatial upsampling convolutions to restore the original image resolution.

Evaluation metrics. We use *forward* error [8, 9], *backward* error [9], and Normalised Mean-squared Error (NME), normalized by inter-ocular distance to report the performance. We train a linear regressor, that maps the discovered landmarks into the ground truth annotations, using a variable number of images in the training set. The learned regressor is tested on the corresponding test partition. Following [8, 9], we refer to this as *forward* error. In addition, [9] also introduced a *backward* error, that trains a regressor in an opposite direction. It maps the ground truth annotations into the discovered landmarks. We use Normalised Mean-squared

© 2023. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

Datasets	AFLW		MAFL	
Methods	F	B	F	B
Sanchez[9]	6.69	10.02	3.99	3.97
Sanchez[9]+Ours	**6.29**	**8.44**	**3.56**	**3.76**

Table 1: Our method is capable of boosting the performance of another competitive baseline [9].

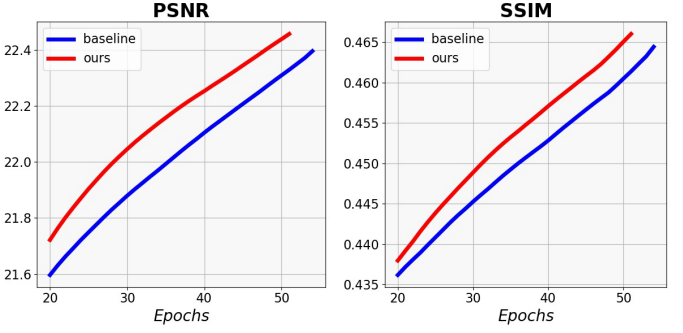

Figure 1: Cumulative PNSR and SSIM [10] over training (on Cats Head) to compare the reconstruction quality between our method and the baseline [9].

Error (NME), normalized by inter-ocular distance to report the performance.

More qualitative results. Figs. 2 and 3 draw additional qualitative comparisons accompanying Sec. 4 (in main paper) on AFLW [6] and MAFL [14] datasets. We see that our method is capable of discovering more semantically relevant landmarks that also capture improved correspondence across different poses and expressions. In contrast, other methods often detect semantically irrelevant landmarks that also lack appropriate correspondence across images.

Fig. 4 shows qualitative comparisons in addition to Sec. 4 (in main paper) on LS3D [1] dataset. We can observe that, in contrast to other methods, our approach is able to discover more semantically meaningful under large pose and expression variations and other challenging factors such as occlusions.

Figs. 5 and 6 display additional qualitative comparisons on Cats Head [13] and Shoes [11, 12] datasets, respectively. In Cats Head dataset, in contrast to others, our method recovers semantically richer landmarks (e.g., around eyes and nose) under different appearance, pose and lighting variations.

With another baseline. We chose another competitive baseline using same loss function [9] to evaluate the effectiveness of our proposed consistency-guided bottleneck (CGB). our CGB, also improves [9] in both forward and backward errors (see Tab. 1).

Reconstruction quality comparison. Fig. 1 shows that, compared to baseline [9], our CGB allows improved reconstruction of the input image.

Varying the range of σ. We study the impact on the performance upon varying the range of σ, Eq.(4) main paper, to which it is mapped (Table 3). Constraining the mapped range between $[0.2, 5]$ provides improved results compared to the relatively bigger range of $[0.2, 10]$. A much bigger range probably over dilates σ, which could likely degrade the reconstruction ability.

Different manifestations of σ. We report performance with different manifestations of σ: fixed, randomly sampled, and the modulated via landmark consistency (Table 2). Modulated

Method	NME%
Fixed σ[4]	3.99
Random σ	4.21
σ(Ours)	**3.50**

Table 2: NME% (forward) with different manifestations of σ.

σ	NME%
[0.2, 5]	**3.50**
[0.2, 10]	3.61

Table 3: NME% (forward) with varying the range of σ mapping.

PS_{update}	NME%
5	3.50
10	3.86
20	3.70
40	**3.36**

Table 4: PS_{update} variations.

σ generally provides improved performance among others, thereby showing the effectiveness of favouring consistent landmarks over noisy counterparts during training.

Varying pseudo-supervision update frequency. We analyze performance upon varying the pseudo-supervision update frequency PS_{update} (Table 4).

Limitations. Like other SOTA methods ([4], [5]), our approach also depends on a pre-trained model trained in a supervised way on an object category. Further, the complexity of KNN affinity graph scales rapidly with more data points. As such, this allows learning some structured representation, presumably shared across different object categories, and hence it could be beneficial for unsupervised landmark discovery task.

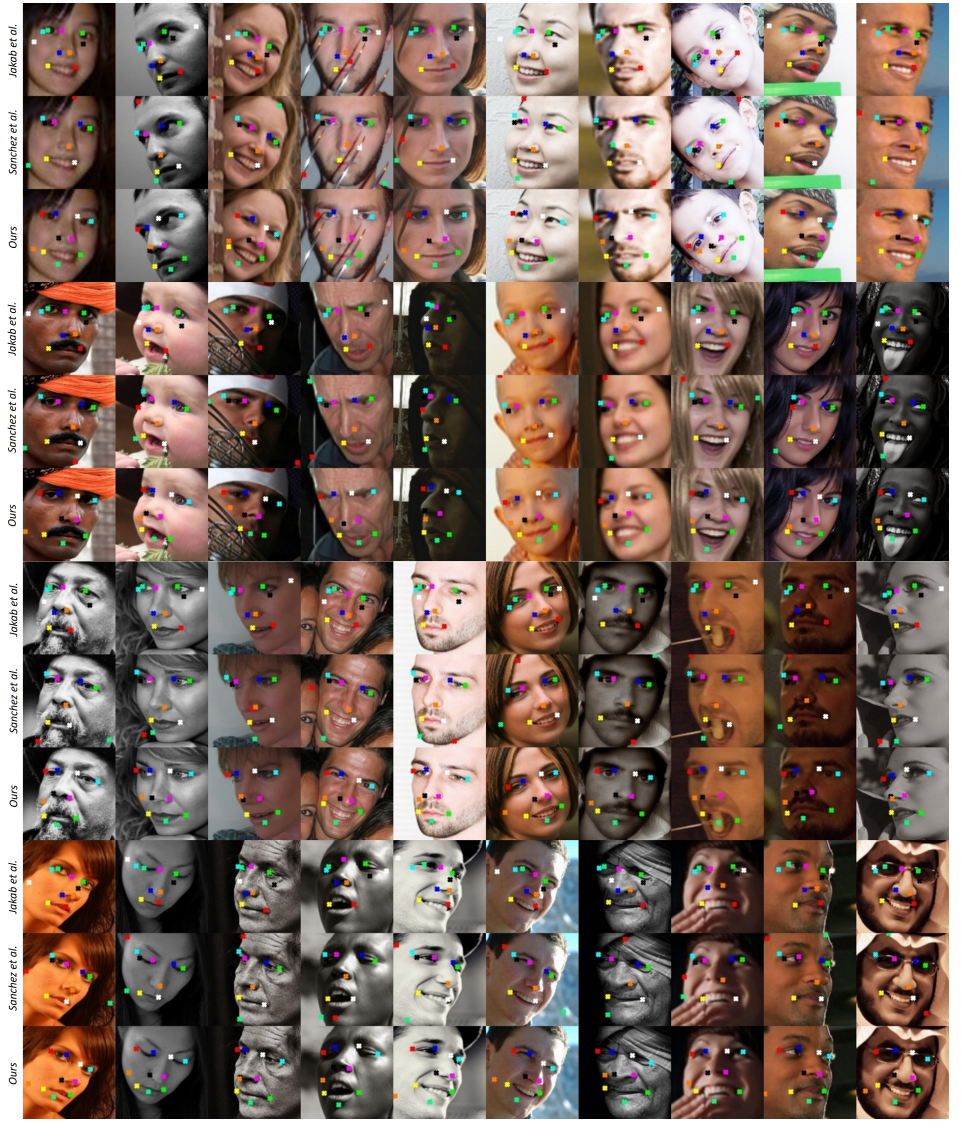

Figure 2: Additional qualitative comparisons on AFLW with Jakab et al. [2](Baseline), and Sanchez et al. [2].

Figure 3: Additional qualitative comparisons on MAFL with Jakab et al. [2](Baseline), and Sanchez et al. [3].

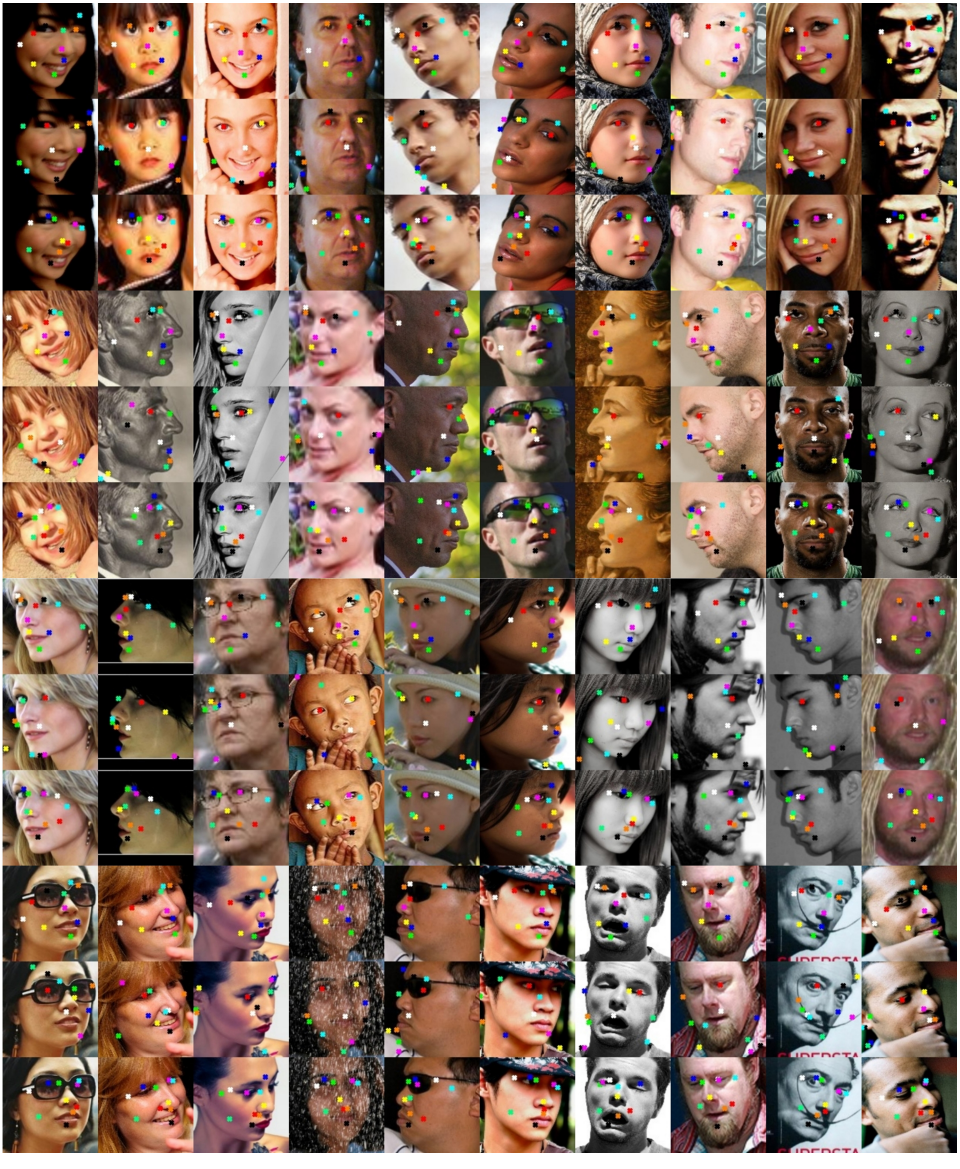

Figure 4: Additional qualitative comparisons on LS3D with Jakab et al. [4](Baseline), and Sanchez et al. [5].

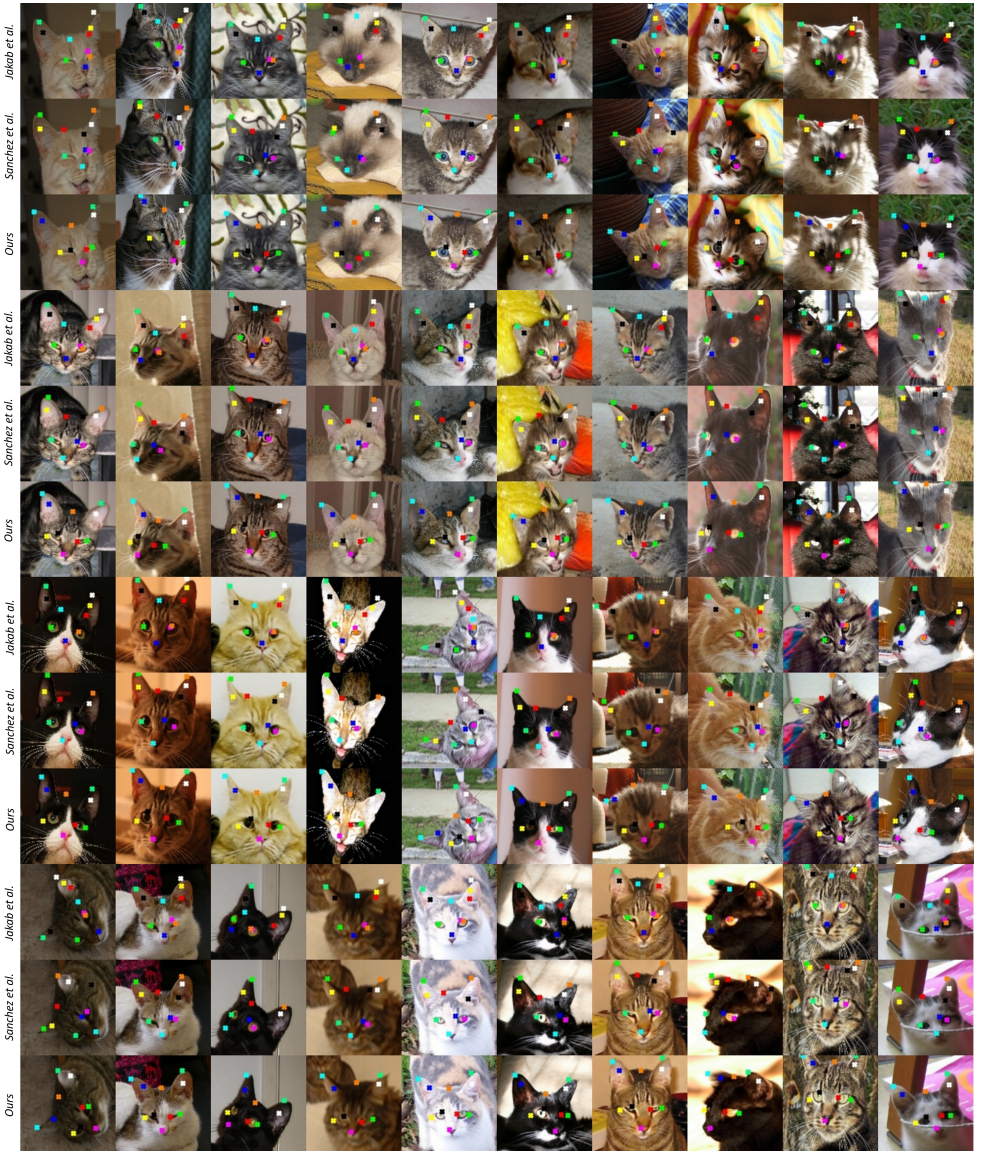

Figure 5: Additional qualitative comparisons on Cats Head with Jakab et al. [4](Baseline), and Sanchez et al. [9].

Figure 6: Additional qualitative comparisons on Shoes with Jakab et al. [4](Baseline), and Sanchez et al. [9].

References

[1] 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.

[2] Yibo Hu, Xiang Wu, Bing Yu, Ran He, and Zhenan Sun. Pose-guided photorealistic face rotation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8398–8406, 2018. doi: 10.1109/CVPR.2018.00876.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017. doi: 10.1109/CVPR.2017.632.

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

[5] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[6] 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.

[7] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. 2018.

[8] 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.

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

[10] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

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

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

[13] 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.

[14] 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.