Fiducial Focus Augmentation for Facial Landmark Detection: Supplementary Material

Purbayan Kar[1]
purbayan.kar@sony.com
Vishal Chudasama[1]
vishal.chudasama1@sony.com
Naoyuki Onoe[1]
naoyuki.onoe@sony.com
Pankaj Wasnik[†1]
pankaj.wasnik@sony.com
Vineeth Balasubramanian[2]
vineethnb@cse.iith.ac.in

[1] Sony Research India, Bangalore, India
[2] Indian Institute of Technology, Hyderabad, India

This supplementary material presents the following details which we could not include in the main paper for space constraints. The additional references are added here.

Contents

† Corresponding Author

S1 Experimental Settings

S1.1 Test Settings

This section describes the test setting configuration we employed in our overall approach. With test images of the datasets (WFLW [1], 300W [28], COFW [5], AFLW [19]), no augmentation - standard augmentations nor Fiducial Focus Augmentation (*FiFA*) - is utilized at test/inference time. With the absence of different augmented views of the input test images, there is no requirement for the Siamese framework (whose purpose is to learn a good feature extractor during training). Consequently, the standalone Transformer + CNN-based backbone is used for testing.

S1.2 Dataset Descriptions

The proposed method was studied on benchmark datasets, namely WFLW [1], 300W [28], COFW [5] and AFLW [19]. Details of these datasets are described below.

Caltech Occluded Faces in-the-Wild (COFW) [5] is a dataset having challenging images with extreme pose variations and occlusion. The dataset comprises 1,345 training and 507 testing images annotated with 29 landmarks.

300 Faces in-the-Wild (300W) [28] is a 68-landmark dataset containing the subsets: AFW [31], LFPW [32], HELEN [33], and XM2VTS [34] with iBUG as an additional dataset. Following common protocol as in earlier work, the provided training splits of HELEN, LFPW and the full set of AFW are used for training, while the test splits of HELEN, LFPW and the iBUG dataset are used for testing. The dataset consists of 3,148 images for training and 689 for testing, containing 554 samples for the common and 135 images for the challenging subsets.

Annotated Facial Landmarks in-the-Wild (AFLW) [19] provides a large-scale collection of annotated facial images sourced from Flickr, exhibiting a large variety of appearance features such as pose, expression, ethnicity, age and gender, alongside general imaging and environmental conditions. The dataset encompasses approximately 25k faces with up to 21 landmarks identified per image.

Wider Facial Landmarks in-the-Wild (WFLW) [1], a recent dataset comprising 98 landmarks, features 7,500 training and 2,500 testing images. In addition to its dense manual annotations, this dataset also incorporates attribute annotations, divided into six subsets: pose, occlusion, expression, blur, make-up, and illumination.

S1.3 Descriptions of Evaluation Metrics

We used different metrics to assess the efficacy of the proposed method in this work. Details of these metrics are described below.

Normalized Mean Error (*NME*) is an extensively used metric to assess the efficacy of a facial landmark localization algorithm. Here, the pixel-wise absolute distance is normalized over a distance that accounts for face size. Then the outcome is obtained by calculating the error of each key point and by averaging it. The NME can be defined mathematically as:

$$NME(P,\bar{P}) = \frac{1}{L}\sum_{l=1}^{L}\frac{||p_l - \bar{p}_l||_2}{d}, \tag{9}$$

where $P, \bar{P}$ denotes the ground truth coordinates of all points and predicted ones for a face image. L is the total number of keypoints, and both $p_l, \bar{p}_l$ are 2-dimensional vectors presenting the x-y coordinates of the i^{th} keypoint. d is the normalization factor denoting the inter-pupil distance or inter-ocular distance (NME_{ic}). The latter could be a distance between the inner corners of the eyes (not commonly used) or the outer corner of the eyes, which is commonly used and used also in our evaluation. Another variant, NME_{box} is computed as the geometric mean of the ground truth bounding box, where $NME_{box} = \sqrt{w_{bbox} \cdot h_{bbox}}$ and NME_{diag} is defined as the diagonal of the bounding box. Following common practice, we use the NME_{ic} for the 300W, WFLW and COFW dataset, while the NME_{box} and NME_{diag} for the AFLW dataset. Here, lower NME indicates better performance.

Failure Rate (FR) offers further comprehension in the configuration of a facial landmark detection algorithm. The NME of each image is evaluated by setting a threshold. Images with an NME surpassing the threshold are classified as failures. The FR is then deduced by evaluating the rate of failures in a given testset. A lower FR indicates better performance. For e.g., FR_{ic}^{10} is computed when corresponding NME_{ic} is calculated as inter-ocular distance and the threshold is set to 10%.

Area Under the Curve (AUC) is another metric widely used by researchers for facial landmark detection. It is obtained through the Cumulative Errors Distribution (CED) curve, plotted from zero to the FR threshold. It results in a non-negative curve whose area is computed to yield the AUC value. An increase in AUC indicates an improvement in the accuracy of predictions for more samples in the test set. AUC_{ic}^{10} is computed when corresponding FR_{ic}^{10} is calculated on the basis of inter-ocular distance. We used AUC_{box} as a metric for the AFLW dataset, where the corresponding FR is calculated on the basis of NME_{box} where NME is computed as the geometric mean of the ground truth bounding box.

S2 Ablation Studies & Analysis

In addition to the ablation studies outlined in the main manuscript, here we present additional ablation experiments to study the efficacy of the proposed method. These experiments involve increasing fiducial patch size, analyzing the anti-aliased hourglass module as well as studying the use of different backbone networks.

S2.1 Increasing Fiducial Patch Size

In order to study whether decreasing the fiducial path size over the training iterations in *FiFA* is the appropriate strategy, we performed experiments reversing the operation to see its effect on overall detection performance. This process, which we term Reversed Fiducial Focus Augmentation (*RFiFA*), involves commencing network training without the use of patches for a specific epoch interval, before gradually introducing 1×1 patches over the landmarks and scaling up to $n \times n$ patches for each subsequent epoch interval. Through our analysis, as presented in Table S1, we determined that an ending patch size of 5×5 yields an NME_{ic} of 3.05. The performance degrades if the ending patch size is either increased or decreased. However, our original proposed *FiFA*, which starts with a 5×5 patch size and gradually reduces it until it is removed from face images, yields the best NME_{ic} of 2.96, validating the proposed *FiFA* strategy.

Table S1: Effect of patch sizes in *FiFA* on COFW.

FiFA patch progression	$NME_{ic}(\%)\downarrow$
no patch $\rightarrow 1\times1 \rightarrow \cdots \rightarrow 3\times3$	3.11
no patch $\rightarrow 1\times1 \rightarrow \cdots \rightarrow 4\times4$	3.08
no patch $\rightarrow 1\times1 \rightarrow \cdots \rightarrow 5\times5$	3.05
no patch $\rightarrow 1\times1 \rightarrow \cdots \rightarrow 6\times6$	3.07
no patch $\rightarrow 1\times1 \rightarrow \cdots \rightarrow 7\times7$	3.09
$5\times5 \rightarrow \cdots \rightarrow 1\times1 \rightarrow$ no patch (Proposed)	**2.96**

Table S2: Effect of different backbone networks in *FiFA* on COFW.

Backbone network	$NME_{ic}(\%)\downarrow$
ResNet-50 [13]	4.02
HRNet [35]	3.31
ViT-B/16 [15]	**3.11**

Table S3: Effect of hourglass components in *FiFA* on COFW.

Method	$NME_{ic}(\%)\downarrow$
Baseline	3.11
+ CNN-based hourglass	3.09
+ anti-aliased CNN-based hourglass	**3.07**

S2.2 Effect of Different Backbone Networks

In our proposed framework, we choose the base variant of Vision Transformer (i.e., ViT-B/16) as our backbone and made necessary modifications to enhance detection performance. To validate our approach, we conducted several experiments by integrating other CNN-based backbones, such as ResNet-50 and HRNet, into our framework. The outcomes of these experiments are presented in Table S2, where it is evident that ViT-B/16 outperforms all other backbones.

S2.3 Effect of Hourglass Components

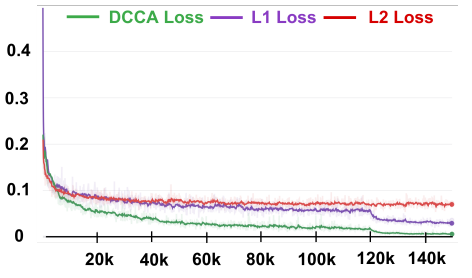
To address the translation variance of CNNs and prevent the loss of structural information, we integrated the anti-aliasing CNN-based hourglass module inside the ViT. To evaluate the efficacy of this anti-aliasing component, we carried out experiments studying the performance of the hourglass network with and without anti-aliased CNNs. The results obtained from these experiments are shown in Table S3 which demonstrates that the anti-aliased hourglass outperforms the simple hourglass model.

S2.4 Effect of loss functions

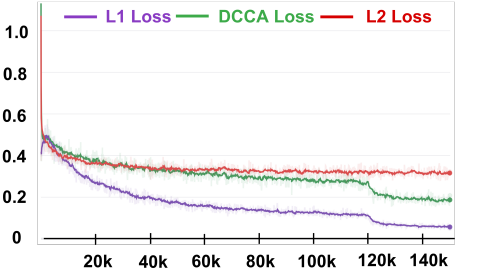
In the main paper, ablation analysis is presented in Table 6 to compare different loss functions. Additionally, in Figure S1, we present different loss curves for further analysis. Figure S1(a) shows the plots for DCCA, $L1$ and $L2$ losses obtained from two representations of Siamese network. From figure, we can see that the DCCA loss converges faster than $L1$ and $L2$ losses. Similarly, Figure S1(b) provides the plots when we replace $L1$ loss between actual and predicted landmark by DCCA and $L2$ losses. One can clearly see that the $L1$ loss converges faster than DCCA and $L2$ losses. These results indicate that in both the scenarios, combination of $L1$ and DCCA loss leads to faster convergence.

S3 Experimental Results

In addition to the experimental analysis presented in the main manuscript, we elaborate on additional experimental results herein. We include an analysis of the *AUC* metric on the



(a) DCCA loss between two representations of Siamese network replaced by $L1$ and $L2$ losses. Furthermore, $L1$ is also employed as a loss between actual and predicted landmark.



(b) $L1$ loss between actual and predicted landmarks replaced by DCCA and $L2$ losses. Furthermore, DCCA is also employed as a loss between two representations of Siamese network.

Figure S1: Comparison between loss functions.

Table S4: Comparison against the state-of-the-art on WFLW testset. Best result is in **bold** and second best result is underlined.

Metric	Models	Remarks	Fullset	Subset					
				Pose	Expression	Illumination	Make Up	Occlusion	Blur
$AUC_{ic}^{10}(\%)\uparrow$	FaRL [43]	$CVPR_{22}$	0.6116	—	—	—	—	—	—
	ADNet [35]	$ICCV_{21}$	0.6002	0.3441	0.5234	0.5805	0.6007	0.5295	0.5480
	SH-FAN [1]	$BMVC_{21}$	**0.6310**	—	—	—	—	—	—
	PropNet [14]	$CVPR_{20}$	0.6158	**0.3823**	**0.6281**	0.6164	**0.6389**	**0.5721**	**0.5836**
	HIH [21]	$ICCVW_{21}$	0.6050	0.3580	0.6010	0.6130	0.6180	0.5390	0.5610
	SLPT [40]	$CVPR_{22}$	0.5950	0.3480	0.5740	0.6010	0.6050	0.5150	0.5350
	DTLD [29]	$CVPR_{22}$	—	—	—	—	—	—	—
	PicassoNet [33]	$TNNLS_{22}$	0.5540	0.2550	0.5100	0.5540	0.5560	0.4600	0.4860
	FiFA (Ours)	—	0.6178	0.3682	0.6024	**0.6219**	0.6255	0.5430	0.5617

WFLW testing dataset through a comparison with state-of-the-art (SOTA) methods, qualitative analysis on 300W, COFW, AFLW datasets, and examples of some failure cases on the test sets.

S3.1 AUC_{ic}^{10} Analysis on WFLW Dataset

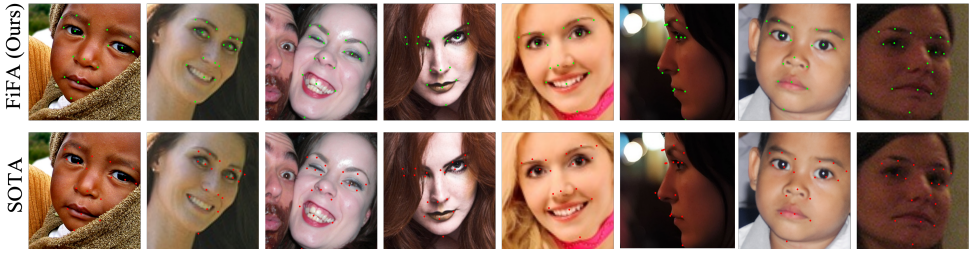
In this study, we compare the proposed method against state-of-the-art methods in terms of AUC_{ic}^{10} for the WFLW test set and its subsets, where a higher AUC_{ic}^{10} indicates better landmark detection performance. Table S4 reveals that while we attain the second-highest AUC_{ic}^{10} for the Fullset, we lag marginally behind [14] for the Pose, Expression, Make Up, Occlusion, and Blur subsets. However, for the Illumination subset, we achieve the highest AUC_{ic}^{10} among other SOTA methods. This signifies that we can predict facial landmarks in a larger fraction of images in the WFLW test set.

S3.2 Additional Qualitative Results

In the main manuscript, we presented a qualitative analysis of the WFLW dataset in comparison with the state-of-the-art method [45]. We herein show qualitative results on other test sets from 300W, COFW and AFLW datasets and compare them with [45]. Our observations, as depicted in Figure S2, indicate that our method outperforms the SOTA approach by delivering more accurate results, particularly in challenging scenarios.

(a) 300W

(b) COFW



(c) AFLW

Figure S2: **Qualitative results on 300W, COFW and AFLW testset**. Landmarks shown in green are produced by our method, while the ones in red by the SOTA approach [45]. Zoom-in for better view.

S3.3 Failure Case Analysis

For completeness of analysis, we additionally present a summary of failure cases of our model, which can promote future work in improving our method. Although our model shows a strong superiority in point of landmark detection, it can be weak on facial images with strong occlusions, particularly those obscured by others, as depicted in Figure S3. Specifically, our model may encounter failure under the following conditions: 1) when challenges such as blurring or occlusion result in significant uncertainty in face-bound inference, and 2) when the face to be aligned is covered by another face, leading to difficulties in distinguishing the target character and resulting in substantial errors. Additionally, ambiguity in landmark annotations can negatively impact the model's performance, particularly for landmarks located at face boundaries. As a potential solution to these limitations, we recommend exploring improved utilization of connections between landmarks to infer the invisible part. This avenue for improvement provides scope for future work.

Figure S3: Examples of failure cases on WFLW, 300W, COFW and AFLW datasets. Blue denotes the ground truth and green represents our predictions. Combinations of low-resolution images, extreme poses, partly covered and overlapping faces make up the majority of failure cases. Zoom-in for better view.

References

[R1] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In 2012 *IEEE conference on computer vision and pattern recognition*, pages 2879–2886. IEEE, 2012.

[R2] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence,* 35(12):2930–2940, 2013

[R3] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European conference on computer vision,* pages 679–692. Springer, 2012

[R4] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, Gilbert Maitre, et al. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication,* volume 964, pages 965–966. Citeseer, 1999