Fiducial Focus Augmentation for Facial Landmark Detection

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

Abstract

Deep learning methods have led to significant improvements in the performance on the facial landmark detection (FLD) task. However, detecting landmarks in challenging settings, such as head pose changes, exaggerated expressions, or uneven illumination, continue to remain a challenge due to high variability and insufficient samples. This inadequacy can be attributed to the model's inability to effectively acquire appropriate facial structure information from the input images. To address this, we propose a novel image augmentation technique specifically designed for the FLD task to enhance the model's understanding of facial structures. To effectively utilize the newly proposed augmentation technique, we employ a Siamese architecture-based training mechanism with a Deep Canonical Correlation Analysis (DCCA)-based loss to achieve collective learning of high-level feature representations from two different views of the input images. Furthermore, we employ a Transformer + CNN-based network with a custom hourglass module as the robust backbone for the Siamese framework. Extensive experiments show that our approach outperforms multiple state-of-the-art approaches across various benchmark datasets.

1 Introduction

Facial Landmark Detection (FLD) aims to detect coordinates of the predefined landmarks on given facial image. The rich geometric information provided by landmarks with distinct semantic significance, such as eye corner, nose tip, or jawline, can be helpful in various tasks like 3D face reconstruction [16, 17, 27], face identification [25, 31, 42], emotion recognition [9, 22, 34], and face morphing [12]. Several FLD algorithms, based either on coordinate regression [24, 30, 32, 33, 43, 46] or heatmap regression [4, 14, 20, 21, 38, 45], have emerged

† Corresponding Author

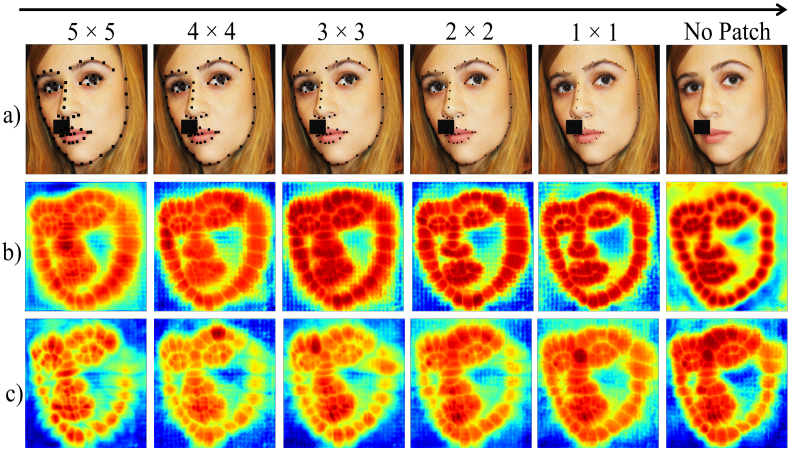

Figure 1: **Illustration of the proposed Fiducial Focus Augmentation (*FiFA*).** In row (a), 5×5 black patches are created around the landmark joints (along with other standard augmentations) in the initial epochs and reduced over the epochs. Rows (b) and (c) show corresponding GradCAM-based saliency maps of the network's last layer with and without *FiFA*, respectively. It is clearly seen that activations are more prominent around the desired landmarks when *FiFA* is used as additional augmentation.

in recent years with promising performance on various datasets. However, landmark detection still remains challenging task due the high variability in poses, lighting and expressions. Despite the various existing FLD methodologies, none have focused on robust image augmentation techniques to solve these challenges. This study illustrates that meticulously designed image augmentations can considerably enhance the FLD performance.

But why do sophisticated deep neural network (DNN) architectures struggle to detect landmarks accurately in challenging scenarios? The reason is that the DNN is unable to learn the facial structure information as accurately as required. If a DNN model can accurately capture features that extract a facial structure, it can predict the landmarks more accurately even from obscured facial regions, like occluded areas. To learn facial structures effectively, we propose new augmentation technique called Fiducial Focus Augmentation (*FiFA*), which leverages the ground truth landmark coordinates as an inductive bias for facial structure. To this end, we introduce $n \times n$ black patches around the landmark locations in the training images, gradually reducing them over the epoch and then removing completely for the rest of the training, as illustrated in Fig 1. Since the patches cover key semantic regions of the face, e.g., eyes, nose, lips and jawline, when the model learns to predict these patches, it is able to learn the entire facial structure significantly better, as compared to an architecture without this inductive bias. One could view this augmentation technique as similar to Curriculum Learning (CL) [1], a strategy that trains a machine learning model from simpler data to more difficult data, mimicking the meaningful order found in human-designed learning curricula.

Drawing inspiration from [2], we leverage the Siamese architecture to acquire a comprehensive understanding of reliable landmark predictions across various image augmentations. However, our method employs Deep Canonical Correlation Analysis (DCCA) [2] as loss function in Siamese architecture to amplify the efficacy of the learning process between distinctively augmented views. This loss function assists in the extraction of features that are correlated across views, while simultaneously eliminating uncorrelated noise. To design a robust backbone for the Siamese architecture, we adopt Vision Transformer (ViT) [8].

We further improved its performance and efficiency by incorporating a Convolutional Neural Network (CNN)-based hourglass module in-between the transformer layers of the ViT. Modern CNNs are usually considered to be shift-invariant; we hence use an Anti-aliased CNN [44] inside the hourglass module to leverage this benefit. We summarize the contributions of this paper as follows.

- To the best of our knowledge, this is the first effort in literature to propose a new patch-based augmentation technique for FLD task to learn facial semantic structures effectively.
- We employ a Siamese-based training scheme utilising DCCA loss between feature representations of two different views of the same image, that enforces consistent predictions of the landmark for the two views. To incorporate virtues of both a Transformer and a CNN, we design a robust Transformer + CNN-based backbone in our proposed framework.
- We performed extensive experiments on various benchmark datasets showing significant improvements over prior work. We also conducted ablation studies on our framework components and additional empirical analysis to study the usefulness of the proposed method.

2 Related Works

Earlier efforts on FLD task, especially those in recent years, can broadly be categorized into network architecture enhancements for heatmap generation and loss function improvements.

Network architecture enhancements: Coordinate regression-based methods [24, 30, 32, 33, 43, 46] directly perform regression on landmark coordinate vectors through a fully connected output layer that disregards the spatial correlations of features and results in limited accuracy of landmark detection. On the other hand, heatmap regression-based methods [3, 4, 14, 15, 20, 21, 38, 40, 45] predict landmark coordinates by creating heatmaps. By doing so, they effectively maintain the original spatial relationships between pixels and achieve promising landmark detection accuracy. Therefore, heatmap regression has become the de facto choice for the FLD task in modern times. In [3], Bulat *et al.* proposed an encoder-decoder based framework with heatmap regression for FLD. Their network incorporates hourglass and hierarchical blocks. Several research works [29, 35, 41] have been published based on the ResNet [13] architecture and modify their network for dense pixel-wise landmark predictions. Recently, the Vision Transformer (ViT) [8] has been incorporated in FLD task by Zhang *et al.* [45] and has produced remarkable results. In our proposed framework, we also use ViT as the backbone network and improve its performance by introducing CNN layers in between transformer layers. This allows us to combine the best of both designs.

Loss function improvements: A pixel-wise $L2$ or $L1$ loss is the conventional loss generally applied to heatmap regression-based methods [6, 7, 26, 37, 47]. To emphasize the importance of tiny and medium range errors during the training process, Feng *et al*. [10] introduced the Wing loss, which modifies the L1 loss by using a logarithmic function to amplify the impact of errors within a specific range. Additionally, Wang *et al*. [36] developed the Adaptive Wing Loss, which can adjust its curvature based on the ground truth pixels. In [18], Kumar *et al*. proposed the LUVLi loss that optimizes the position of the keypoints, the uncertainty, and the likelihood of visibility. Recently, the authors from [14] proposed the Focal Wing Loss, which is used to mine and emphasize difficult samples under in-the-wild conditions.

In this work, we use the standard Binary Cross Entropy (BCE) and $L2$ losses for heatmap and coordinate regression, respectively. We however employ the DCCA loss [2] which suits our framework and has never been used before for the FLD task. These simple losses help

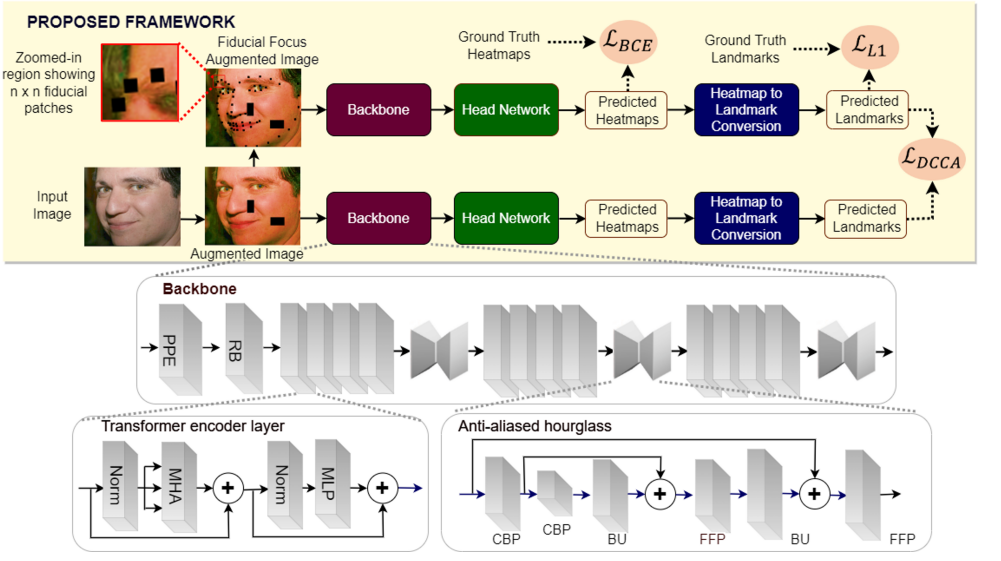

Figure 2: **An overview of the proposed Siamese-based framework**. PPE = Patch + Position Embeddings; RB = Residual Block; MHA = Multi-Head Attention, MLP = Multi-Layer Perceptron; CBP = Convolution+BlurPool; BU = Bilinear Upsampling; FFP = FF-Parser.

the proposed framework set a new benchmark. Our study of literature revealed that well-designed image augmentations are largely ignored for the FLD task. This paper attends to this very issue and introduces a new augmentation technique called *FiFA* that accounts for our impressive results.

3 Proposed Framework

3.1 Problem Statement & Notations

Given an input image I, FLD aims to detect $\{x,y\} \in \mathbb{R}^{k\times 2}$, the coordinates of K predefined landmarks. To this end, we propose a heatmap-based approach to regress the facial landmarks. During training, it encodes the target ground truth coordinates as a series of k heatmaps with a 2D Gaussian curve centered on them:

$$\Psi_{i,j,k} = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}[(i-\bar{x}_k)^2+(j-\bar{y}_k)^2]} \tag{1}$$

where x_k and y_k are the spatial coordinates of the k^{th} point, while $\bar{x}_k$ and $\bar{y}_k$ are their scaled, quantized version obtained by scaling factor s and rounding operator $\lfloor\cdot\rceil$, i.e.

$$(\bar{x}_k, \bar{y}_k) = (\lfloor\frac{1}{s}x_k\rceil, \lfloor\frac{1}{s}y_k\rceil) \tag{2}$$

As shown in Eq. (1), we use a Gaussian with variance σ around each coordinate from $\{x,y\}$ to generate the corresponding heatmap $\mathbb{H} \in \mathbb{R}^{k\times W\times H}$. Finally, the pixels with maximum intensity of the heatmap $\mathbb{H}$ are selected to get the final K landmarks in the FLD task.

To attain precise facial landmarks, we propose a novel augmentation technique called Fiducial Focus Augmentation (*FiFA*) that helps the network to learn facial structures in the

provided images, along with a Siamese network with a robust backbone and the DCCA loss to ensure consistent predictions between different augmented views. Detailed explanations of these modules are provided in the subsequent subsections.

3.2 Fiducial Focus Augmentation

We seek to explore the potential of carefully designed image augmentations for the FLD task in this section. To this end, we propose an augmentation f_A for input training images, where $f_A = f_{A_2} \circ f_{A_1}$. Here, f_{A_1} can be any standard image augmentations used in the FLD task [9, 14, 38, 40, 45] and f_{A_2} is the proposed Fiducial Focus Augmentation (*FiFA*).

First, we take the original input image I and apply standard image augmentation f_{A_1} to get the augmented image (I'). Mathematically, this can be expressed as:

$$I' = f_{A_1} \otimes I. \tag{3}$$

To get the final augmented image I'', I' is passed through the proposed augmentation operation i.e., f_{A_2} (as descibed in Alg 1), i.e.

$$I'' = f_{A_2} \otimes I' = \hat{I} \otimes I' = \hat{I} \otimes (f_{A_1} \otimes I). \tag{4}$$

Here, we aim to incorporate the available facial structure ground truth information into the augmented image, I' in order to aptly utilize the underlying facial structure. To achieve this, we construct black square patches of dimensions $h_f \times w_f$, where $h_f, w_f \in \{1, \cdots, n\}$ while retaining the landmarks as the intersection points of the two diagonals of the square patches (see Figure 1 (a)). These patches comprise of four coordinates which can be expressed as:

$$\{(x_i - w_f, y_i + h_f), (x_i + w_f, y_i + h_f), (x_i + w_f, y_i - h_f), (x_i - w_f, y_i - h_f)\} \; \forall \; \{x_i, y_i\} \in L. \tag{5}$$

Here, we start with a bigger patch size of $n \times n$ for a certain number of epoch intervals $\mathcal{E}$. After every such interval, we reduce the patch size by 1 pixel and eventually, these patches are removed from the images and rest of the training goes on with augmentation f_{A_1} only. So the final augmented image is (where T_n is the total number of epochs):

$$\begin{cases} I'' & \text{when epoch no.} \leq n \cdot \mathcal{E} \\ I' & \text{when } n \cdot \mathcal{E} < \text{epoch no.} \leq T_n. \end{cases} \tag{6}$$

Algorithm 1 Fiducial Focus Augmentation (f_{A_2})

Initialize: I': Augmented Image, where $I' = f_{A_1} \otimes I$,
L_n: Number of landmarks in I,
L: Set of L_n landmarks, where $L = \{(x_i, y_i)\}$, $i \in \{1, ..., L_n\}$,
h_f, w_f: Height and width of the patches (S), where $h_f, w_f \in \{n, ..., 1\}$,
I_{in}: Pixel intensity of S, where $I_{in} = (0, 0, 0)$,
$\mathcal{E}$: Epoch interval, where $\mathcal{E} \in \{1, ..., n\} \wedge n <$ Total number of epochs (T_n)
T_n : Total number of epochs $= \sum_{i=1}^{n} \mathcal{E}_i + w$, where $w \in \mathbb{W}$
Procedure:
  for i in range T_n **do**
    k $\leftarrow \lfloor i/\mathcal{E} \rfloor$
    $h_f, w_f \leftarrow |n - k|$
    for j in range L_n **do**
      $C \leftarrow \{(x_j - w_f/2, y_j + h_f/2), (x_j + w_f/2, y_j + h_f/2), (x_j + w_f/2, y_j - h_f/2), (x_j - w_f/2, y_j - h_f/2)\}$
      Create patch S with C of I_{in}
      $I' \leftarrow S \otimes I'$
    end for
  end for
  $\hat{I} \leftarrow I'$
  return $\hat{I}$

The proposed *FiFA* helps the backbone network learn the underlying facial structure and address difficult test samples, since the patches cover the entire face uniformly over the different joints (eyes, lips, nose and jawline). At the beginning of training, the model is exposed to larger patches as low-confidence regions to concentrate on the joints and eventually, as the model learns progressively with each epoch, smaller patches are introduced as high-confidence regions around the joints. When the patches are removed completely, the model tries to predict the joints with the inductive bias provided by earlier training steps in our augmentation process. Since the patches can be used with any facial variations (such as pose or expression), their integration into the images as augmentations enables the model to learn the inherent facial structures.

3.3 Matching Two Views

Earlier work on the task of FLD has seen limited exploration of Siamese architecture-based training, with the exception of [4]. In this paper, we propose a Siamese architecture-based framework as illustrated in Fig. 2. The network f takes the two input images I' and I'' generated using two different augmentations f_{A_1} and f_A. This training scheme using augmentations holds a notable advantage, as CNNs may not be invariant under arbitrary affine transformations. Therefore, even minor variations within the input space may produce significant changes in the output. By optimizing jointly using the Siamese architecture and combining the two predictions, we enhance the robustness and consistency of the predictions (under such variations).

To maximize the correlation between two different augmented views, we employ the Deep Canonical Correlation Analysis (DCCA) loss [2] between the high-level representation mappings $f_1(I')$ and $f_2(I'')$, where $f_1 = f_2 = f$. The correlation between these two mappings can be expressed as below:

$$corr(f_1(I'), f_2(I'')) = \frac{cov(f_1(I'), f_2(I''))}{\sqrt{var(f_1(I')) \cdot var(f_2(I''))}}. \tag{7}$$

The DCCA loss (i.e., $\mathcal{L}_{DCCA}$) is then computed as:

$$\mathcal{L}_{DCCA} = -corr(f_1(I'), f_2(I'')). \tag{8}$$

The use of DCCA loss presents three key advantages: (i) correlated representations partially reconstruct the information in the second view, when it is unavailable; (ii) it has potential to eliminate noise that is uncorrelated across the two views; and (iii) if f_1, f_2 capture features that are correlated across the views, they may represent latent aspects of the face. This, in turn helps the backbone network in capturing the facial structure in the images.

3.4 Architectural Details

In the proposed framework, we employ a transformer-based architecture (a pre-trained ViT-B/16 [45] consisting of 12 layers and a width of 768) as a backbone. To enhance its performance further, we incorporated three custom CNN-based hourglass modules after every four layers of the transformer network. The purpose of this module is to introduce desirable properties of CNNs, such as shift, scale, and distortion invariance, into the ViT architecture, while still retaining the characteristics of transformers, i.e., dynamic attention, global context, and better generalization. This results in a robust backbone network (Transformer + CNN) which learns facial structures effectively.

The utilization of pooling layers in CNNs often provides a certain degree of shift invariance in the model. However, in our task, it is imperative to avoid the loss of structural information caused by pooling layers. We therefore adopt the Anti-aliased CNN [44] into our hourglass modules, hereafter known as Anti-aliased Hourglass. The combination of these components significantly enhances the caliber of our network towards high-quality heatmap generation. Nevertheless, the upsampling + concatenation (U+A) operation in the hourglass modules may introduce some high-frequency noise. To mitigate this negative impact and filter the features in the Fourier space, we integrate a FF-Parser layer [39] after each U+A operation in the hourglass modules. We provide ablation studies on these components in our results to demonstrate their usefulness.

4 Experiments and Results

This section discusses the implementation details, comparison with SOTA methods on benchmark datasets and ablation analysis of the introduced components of the proposed method.

Implementation Details: The proposed method is trained/tested on the various benchmark datasets, i.e., WFLW [1], 300W [28], COFW [6] and AFLW [19]. Details of these datasets are discussed in the Supplementary material. During the training phase, the input image is cropped and resized to 512×512. The output feature map size of every hourglass module is set to 128×128, which is $4\times$ smaller than the input image size. The ground truth heatmaps are generated by a Gaussian with $\sigma = 1.5$ and radius $r = 5$. During training process, we used AdamW [23] to optimize our network with the initial learning rate of 1×10^{-4} and trained for 250 epochs. Apart from the proposed augmentation (*FiFA*), other standard data augmentations (f_{A_1}) are employed at training time, such as random masking, bilinear interpolation, random occlusion, random gray, random gamma, random blur, noise fusion. For effective learning, along with the DCCA loss (i.e., $\mathcal{L}_{DCCA}$), we also employ the standard BCE loss (i.e., $\mathcal{L}_{BCE}$) and mean absolute error loss (i.e., $\mathcal{L}_{L1}$) for heatmap and coordinate regression, respectively with equal weights (i.e., 1.0). For evaluation, we used the standard evaluation metrics i.e., Normalized Mean Error (NME) variants (i.e., NME_{ic}, NME_{box}, NME_{diag}), Failure Rate (FR^{10}_{ic}), Area Under the Curve (AUC_{box}). Detailed definitions of these metrics have been discussed in the Supplementary material. For comparison, we choose recent baselines such as FaRL [45], ADNet [15], SH-FAN [4], PropNet [14], HIH [20], SLPT [40], PicassoNet [38] and DTLD [21]. All the experiments were implemented using PyTorch and the network was trained on 4 GPUs (40GB NVIDIA A100), with batch size 5 per GPU.

4.1 Result Analysis

Comparison on COFW: In Table 1, we presents a comparison of the proposed *FiFA* approach with existing SOTA methods on the COFW testset, which is a well-known benchmark for heavy occlusion and a wide range of head pose variation. It is noteworthy that the proposed *FiFA* model outperforms the existing SOTA methods. The leading NME_{ic} and 0% FR^{10}_{ic} demonstrate its robustness against extreme situations.

Comparison on 300W: On the 300W dataset, our approach exhibits superior performance in comparison to SOTA methods in terms of NME_{ic}, and is given in Table 1. In challenge-set, the proposed approach performs slightly lower than PropNet [14] and SH-FAN [4] methods. However, it has achieved SOTA results in other scenarios (i.e., full-set and common-set), which suggests that our method makes plausible predictions even in deplorable situations.

Table 1: Comparison against the state-of-the-art on COFW, 300W and AFLW dataset. Best result is **bolded** and second best result is underlined.

Method	Remarks	COFW		300W			AFLW			
		$NME_{ic}\downarrow$	$FR_{ic}^{10}\downarrow$	$NME_{ic}\downarrow$			$NME_{diag}\downarrow$		$NME_{box}\downarrow$	$AUC_{box}\uparrow$
				Full	Common	Challenge	Full	Frontal	Full	Full
FaRL [45]	$CVPR_{22}$	3.11	0.12	2.93	2.56	4.45	0.94	0.82	1.33	81.3
ADNet [35]	$ICCV_{21}$	4.68	0.59	2.93	2.53	4.58	—	—	—	—
SH-FAN [40]	$BMVC_{21}$	3.02	**0.00**	2.94	2.61	4.13	1.31	1.12	2.14	70.0
PropNet [24]	$CVPR_{20}$	3.71	0.20	2.93	2.67	**3.99**	—	—	—	—
HIH [20]	$ICCVW_{21}$	3.21	**0.00**	3.09	2.65	4.89	—	—	—	—
SLPT [44]	$CVPR_{22}$	3.32	0.59	3.17	2.75	4.90	—	—	—	—
DTLD [21]	$CVPR_{22}$	3.02	—	2.96	2.60	4.48	1.37	—	—	—
PicassoNet [33]	$TNNLS_{22}$	—	—	3.58	3.03	5.81	1.59	1.30	—	—
FiFA (Ours)	—	**2.96**	**0.00**	**2.89**	**2.51**	4.47	**0.92**	**0.80**	**1.31**	**81.8**

Table 2: Comparison against the state-of-the-art on WFLW testset. Best result is **bolded** and second best result is underlined.

Metric	Models	Remarks	Fullset	Subset					
				Pose	Expression	Illumination	Make Up	Occlusion	Blur
$NME_{ic}(\%)\downarrow$	FaRL [45]	$CVPR_{22}$	3.99	6.61	4.18	3.90	3.84	4.71	4.53
	ADNet [35]	$ICCV_{21}$	4.14	6.96	4.38	4.09	4.05	5.06	4.79
	SH-FAN [40]	$BMVC_{21}$	**3.72**	—	—	—	—	—	—
	PropNet [24]	$CVPR_{20}$	4.05	6.92	3.87	4.07	**3.76**	**4.58**	**4.36**
	HIH [20]	$ICCVW_{21}$	4.08	6.87	**4.06**	4.34	3.85	4.85	4.66
	SLPT [44]	$CVPR_{22}$	4.14	6.96	4.45	4.05	4.00	5.06	4.79
	DTLD [21]	$CVPR_{22}$	4.05	—	—	—	—	—	—
	PicassoNet [33]	$TNNLS_{22}$	4.82	8.61	5.14	4.73	4.68	5.91	5.56
	FiFA (Ours)	—	3.89	**6.47**	4.09	**3.80**	**3.76**	4.63	4.43
$FR_{ic}^{10}(\%)\downarrow$	FaRL [45]	$CVPR_{22}$	1.76	—	—	—	—	—	—
	ADNet [35]	$ICCV_{21}$	2.72	12.72	2.15	2.44	1.94	5.79	3.54
	SH-FAN [40]	$BMVC_{21}$	**1.55**	—	—	—	—	—	—
	PropNet [24]	$CVPR_{20}$	2.96	12.58	2.55	2.44	1.46	5.16	3.75
	HIH [20]	$ICCVW_{21}$	2.60	12.88	**1.27**	2.43	**1.45**	5.16	3.10
	SLPT [44]	$CVPR_{22}$	2.76	12.72	2.23	1.86	3.40	5.98	3.88
	DTLD [21]	$CVPR_{22}$	2.68	—	—	—	—	—	—
	PicassoNet [33]	$TNNLS_{22}$	5.64	25.46	5.10	4.30	5.34	10.59	7.12
	FiFA (Ours)	—	1.60	**7.05**	**1.27**	**1.43**	**1.45**	**3.39**	**1.94**

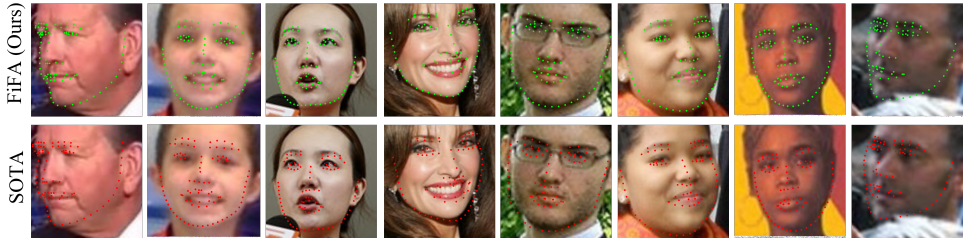

Figure 3: **Qualitative results on WFLW testset**. Landmarks shown in green are produced by our method, while the ones in red by the state-of-the-art approach of [45].

Comparison on AFLW: The results on AFLW testset are presented in Table 1. Adhering to the evaluation protocol adopted in [45], we report comparisons in terms of NME_{diag}, NME_{box} and AUC_{box}^{7}. This table clearly indicates that our approach has outperformed the SOTA results, despite the fact that the dataset is almost saturated.

Comparison on WFLW: In Table 2, we compare results in terms of NME_{ic}, and FR_{ic}^{10}. Here, it is observed that the proposed *FiFA* approach obtains better NME_{ic} for Pose, Illumination and Make Up subsets. Additionally, in comparison on FR_{ic}^{10}, the proposed approach achieves

Table 3: Effect of method's components on COFW.

Method	$NME_{ic}(\%)\downarrow$
Vanilla backbone (ViT-B/16)	3.11
+ anti-aliased CNN-based hourglass	3.07
+ Fiducial Focus Augmentation	3.00
+ Siamese training (w DCCA)	**2.96**

Table 4: Effect of patch sizes in *FiFA* on COFW.

FiFA patch progression	$NME_{ic}(\%)\downarrow$
$3\times3 \rightarrow \cdots \rightarrow 1\times1 \rightarrow$ *no patch*	3.05
$4\times4 \rightarrow \cdots \rightarrow 1\times1 \rightarrow$ *no patch*	3.00
$5\times5 \rightarrow \cdots \rightarrow 1\times1 \rightarrow$ *no patch*	**2.96**
$6\times6 \rightarrow \cdots \rightarrow 1\times1 \rightarrow$ *no patch*	2.99
$7\times7 \rightarrow \cdots \rightarrow 1\times1 \rightarrow$ *no patch*	3.02

Table 5: Effect of *FiFA* over standard augmentations on COFW. BI = Bilinear Interpolation; RM = Random Masking; RO = Random Occlusion; RGr = Random Gray; RGm = Random Gamma; RB = Random Blur; NF = noise fusion.

Augmentations	$NME_{ic}(\%)\downarrow$
RM + RO	3.15
+ *FiFA*	3.08
RM + {RO, RGr}	3.12
+ *FiFA*	3.07
RM + {RO, RGr, RGm}	3.10
+ *FiFA*	3.04
RM + {RO, RGr, RGm, RB}	3.10
+ *FiFA*	3.04
RM + BI + {RO, RGr, RGm, RB}	3.08
+ *FiFA*	3.03
RM + BI + NF + {RO, RGr, RGm, RB}	3.07
+ *FiFA*	3.00

higher performance in all subsets i.e., Pose, Expression, Illumination, Make Up, Occlusion, Blur by 44%, 41%, 23%, 1%, 34%, 37.4%, respectively over the previous best performing SOTA methods. These results show that our method improves the accuracy in challenging scenarios while also reducing the overall failure ratio for difficult images. Moreover, Fig. 3 visually conveys that the proposed approach delivers significantly more precise landmarks in challenging scenarios.

4.2 Ablation Studies & Analysis

This section presents the ablation analysis carried out to establish the efficacy of the proposed framework. To ensure fair comparison, all experiments were performed on COFW dataset.

Effects of method's components: Herein, we investigate the impact of each component of the proposed framework. The results, presented in Table 3, reveal that the baseline network, i.e., Vanilla backbone (ViT-B/16), attains an NME_{ic} of 3.11 when trained solely with standard augmentations, i.e., f_{A_1}. When anti-aliased CNN-based hourglass modules are incorporated into baseline, an improvement in NME_{ic} to 3.07 is observed. By employing the proposed augmentation, f_{A_2}, on the input images during training, a remarkable performance boost is achieved, with an NME_{ic} of 3.00. The highest NME_{ic} of 2.96 is attained when incorporating the Siamese training approach with DCCA loss on both f_{A_1} and f_{A_2} augmented images. This finding demonstrates that training the backbone with proposed components gives best performance in results.

Effects of fiducial mask sizes: We have conducted a series of experiments to determine the optimal initial patch size for the proposed *FiFA*. As shown in Table 4, a patch size of 5×5 yields the best NME_{ic} of 2.96, while deviating from this size leads to a deterioration in performance. This can be attributed to the fact that during the initial stages of training, when the network weights are not yet sufficiently tuned, a patch size that is either too large or too small will result in a confidence region that is either too broad or too narrow for the network to focus on the landmarks. This, in turn, has an adverse effect on the learning process and ultimately on the performance of the network.

Effect of *FiFA* over standard augmentations: Several experiments were conducted to prove the effectiveness of our proposed *FiFA* over other standard augmentations. Due to the

availability of only one view of augmented images, all these experiments were performed without a Siamese-based training mechanism. Table 5 displays the results obtained in terms of NME_{ic} on the COFW testset. One can notice that the inclusion of our proposed *FiFA* in standard augmentation techniques leads to a notable improvement in the NME_{ic} value.

Comparison with other losses in Siamese training: We employ DCCA loss [2] in Siamese training to maximize the correlation between different views. To demonstrate the efficacy of DCCA loss, we conducted several experiments with different losses (i.e., L2, L1, Smooth L1, and Wing loss [10]), and the corresponding results are presented in Table 6. One can observe that the DCCA loss helps to obtain better NME_{ic}, exhibiting a 3% increase as compared to previous best-performing Wing loss.

Table 6: Effect of different losses in Siamese training on COFW.

Loss	L2	L1	Smooth L1	Wing [10]	DCCA [2]
$NME_{ic}(\%)\downarrow$	3.14	3.09	3.11	3.05	**2.96**

Effectiveness of the proposed components to other SOTA methods: To validate the effectiveness of the proposed components, we conducted a series of experiments wherein the proposed *FiFA* augmentation and Siamese network based DCCA loss were implemented on other baseline methods such as HRNet [33], ADNet [15], SH-FAN backbone [4], FaRL [45], SLPT [40] and the corresponding results are summarized in Table 7. The proposed *FiFA* augmentation technique improved the performance of baseline methods. Additionally, the Siamese network based DCCA loss contributed to improve the NME score further. This clearly indicates the generalization capability of our method.

Table 7: Effect of proposed *FiFA* augmentation technique and Siamese-based DCCA loss on baseline methods on COFW testset.

Methods	Remarks	Baseline	+ *FiFA*	+ *FiFA* + Siamese training (w DCCA)
HRNet [33]	$ICCV_{21}$	3.45	3.32	**3.28**
ADNet [15]	$ICCV_{21}$	4.68	4.51	**4.45**
SH-FAN Backbone [4]	$BMVC_{21}$	3.25	3.12	**3.07**
FaRL [45]	$CVPR_{22}$	3.11	3.04	**3.01**
SLPT [40]	$CVPR_{22}$	3.32	3.15	**3.10**

5 Conclusion & Future Work

In this paper, we successfully proposed a simple yet effective image augmentation technique called Fiducial Focus Augmentation (*FiFA*) for facial landmark detection task. The integration of *FiFA* during training significantly enhanced the accuracy of proposed approach on testing benchmarks without extreme modifications to its backbone network and the loss function. Our findings suggest that the employment of *FiFA* as an image augmentation technique, when used in conjunction with a Siamese-based training with DCCA loss results in state-of-the-art performance. Additionally, we employed an anti-aliased CNN-based hourglass network with ViT as our backbone network to address shift invariance and noise. We performed extensive experimentation and ablation studies to validate the effectiveness of the proposed approach. In future work, *FiFA* can be studied further to extend it for other face-related tasks.

References

[1] Look at boundary: A boundary-aware face alignment algorithm. https://wywu.github.io/projects/LAB/WFLW.html.

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013.

[3] 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 international conference on computer vision*, pages 1021–1030, 2017.

[4] Adrian Bulat, Enrique Sanchez, and Georgios Tzimiropoulos. Subpixel heatmap regression for facial landmark localization. *arXiv preprint arXiv:2111.02360*, 2021.

[5] Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *2013 IEEE International Conference on Computer Vision*, pages 1513–1520, 2013. doi: 10.1109/ICCV.2013.191.

[6] Jiankang Deng, George Trigeorgis, Yuxiang Zhou, and Stefanos Zafeiriou. Joint multi-view face alignment in the wild. *IEEE Transactions on Image Processing*, 28(7):3636–3648, 2019.

[7] 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.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570, 2016.

[10] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2235–2245, 2018.

[11] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *International Conference on Machine Learning*, pages 2535–2544. PMLR, 2019.

[12] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4295–4304, 2015.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. cvpr. 2016. *arXiv preprint arXiv:1512.03385*, 2016.

[14] Xiehe Huang, Weihong Deng, Haifeng Shen, Xiubao Zhang, and Jieping Ye. Propagationnet: Propagate points to curve to learn structure information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7265–7274, 2020.

[15] Yangyu Huang, Hao Yang, Chong Li, Jongyoo Kim, and Fangyun Wei. Adnet: Leveraging error-bias towards normal direction in face alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3080–3090, 2021.

[16] Josef Kittler, Patrik Huber, Zhen-Hua Feng, Guosheng Hu, and William Christmas. 3d morphable face models and their applications. In *Articulated Motion and Deformable Objects: 9th International Conference, AMDO 2016, Palma de Mallorca, Spain, July 13-15, 2016, Proceedings 9*, pages 185–206. Springer, 2016.

[17] Paul Koppen, Zhen-Hua Feng, Josef Kittler, Muhammad Awais, William Christmas, Xiao-Jun Wu, and He-Feng Yin. Gaussian mixture 3d morphable face model. *Pattern Recognition*, 74:617–628, 2018.

[18] 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.

[19] Martin Köstinger, 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 International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2144–2151, 2011. doi: 10.1109/ICCVW.2011.6130513.

[20] Xing Lan, Qinghao Hu, Qiang Chen, Jian Xue, and Jian Cheng. Hih: Towards more accurate face alignment via heatmap in heatmap. *arXiv preprint arXiv:2104.03100*, 2021.

[21] Hui Li, Zidong Guo, Seon-Min Rhee, Seungju Han, and Jae-Joon Han. Towards accurate facial landmark detection via cascaded transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4176–4185, 2022.

[22] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017.

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[24] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3317–3326, 2017.

[25] Iacopo Masi, Stephen Rawls, Gérard Medioni, and Prem Natarajan. Pose-aware face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4838–4846, 2016.

[26] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 483–499. Springer, 2016.

[27] Joseph Roth, Yiying Tong, and Xiaoming Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4197–4206, 2016.

[28] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016.

[29] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019.

[30] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[31] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[32] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

[33] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4187, 2016.

[34] Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Copula ordinal regression for joint estimation of facial action unit intensity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4902–4910, 2016.

[35] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

[36] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6971–6981, 2019.

[37] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[38] Tiancheng Wen, Zhonggan Ding, Yongqiang Yao, Yaxiong Wang, and Xueming Qian. Picassonet: Searching adaptive architecture for efficient facial landmark localization. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[39] Junde Wu, Huihui Fang, Yu Zhang, Yehui Yang, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. *arXiv preprint arXiv:2211.00611*, 2022.

[40] Jiahao Xia, Weiwei Qu, Wenjian Huang, Jianguo Zhang, Xi Wang, and Min Xu. Sparse local patch transformer for robust face alignment and landmarks inherent relation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4052–4061, 2022.

[41] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[42] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4362–4371, 2017.

[43] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 1–16. Springer, 2014.

[44] Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019.

[45] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022.

[46] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 386–391, 2013.

[47] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 386–391, 2013.