Fiducial Focus Augmentation for Facial Landmark Detection

Purbayan Kar[1], Vishal Chudasama[1], Naoyuki Onoe[1], Pankaj Wasnik[1], Vineeth Balasubramanian[2]
{purbayan.kar, vishal.chudasama1, naoyuki.onoe, pankaj.wasnik}@sony.com, vineethnb@cse.iith.ac.in
[1]Sony Research India, Bangalore, India [2]Indian Institute of Technology, Hyderabad, India

BMVC 2023

1. Introduction:

- Facial Landmark Detection (FLD) aims to detect coordinates of the predefined landmarks on given facial image.
- Can be helpful in various computer vision applications such as 3D face reconstruction, face identification and emotion recognition.
- FLD is challenging due to high variability in poses, lightning and expressions.
- Existing FLD algorithms are either based on coordinate regression or heatmap regression. None of these methods have focused on robust image augmentation techniques to solve the challenges.

2. Objective:

- This work aims to predict accurate landmarks on faces to learn facial semantic structures effectively and outperforms other state-of-the-art (SOTA) methods.

3. Contributions:

- A new patch-based augmentation technique called Fiducial Focus Augmentation (FiFA) is proposed for FLD task to learn facial semantic structures effectively.
- We employ a Siamese-based training scheme utilizing Deep Canonical Correlation Analysis (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.

4. Methodology:

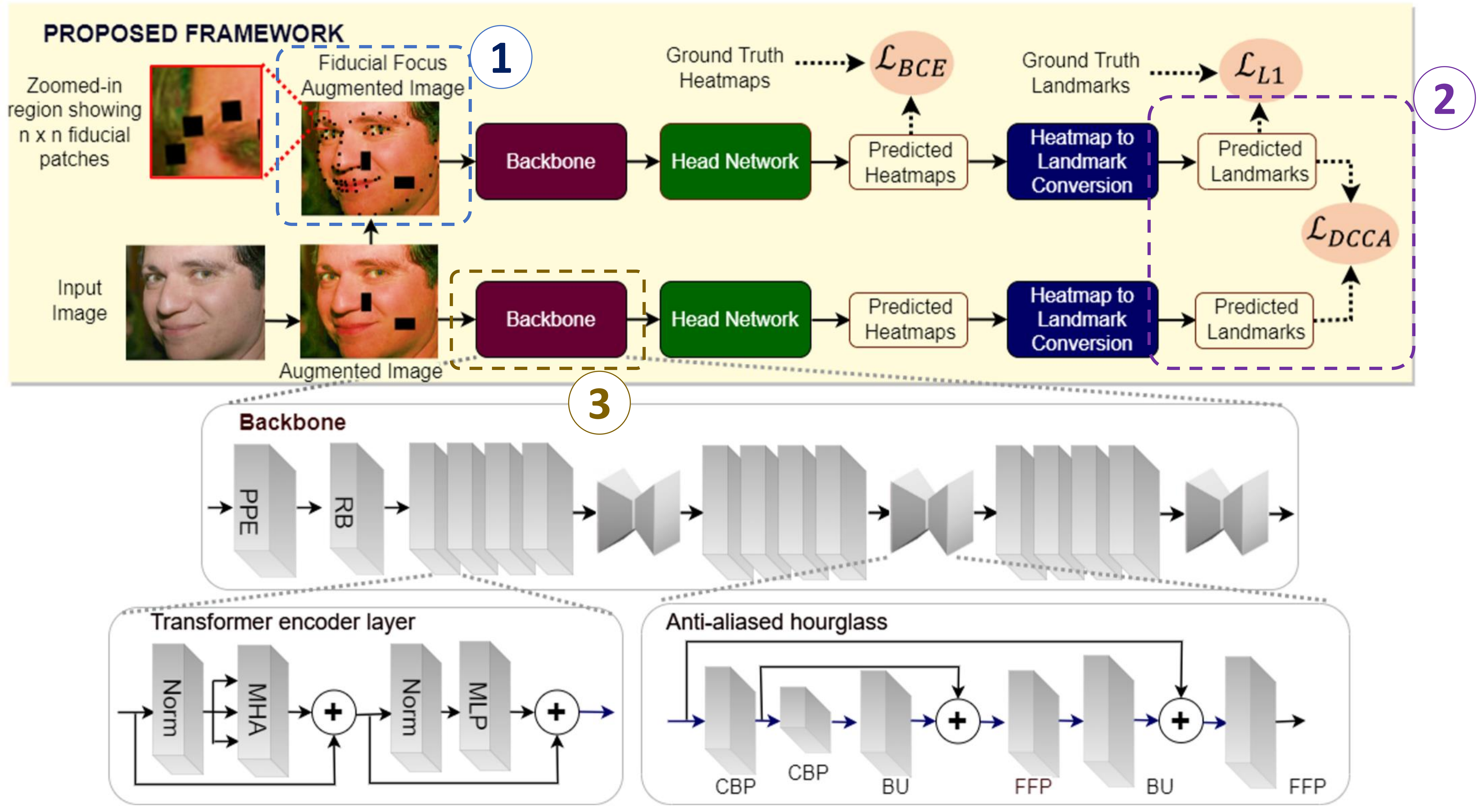

Fig. 1 :An overview of the proposed Siamese-based framework. PPE = Patch + Position Embeddings; RB = Residual Block; MHA = Multi-Head Attention, MLP = Multi-Layer Perceptron; CBP= Convolution+BlurPool; BU = Bilinear Upsampling; FFP = FF-Parser.

1. Fiducial Focus Augmentation (*FiFA*)

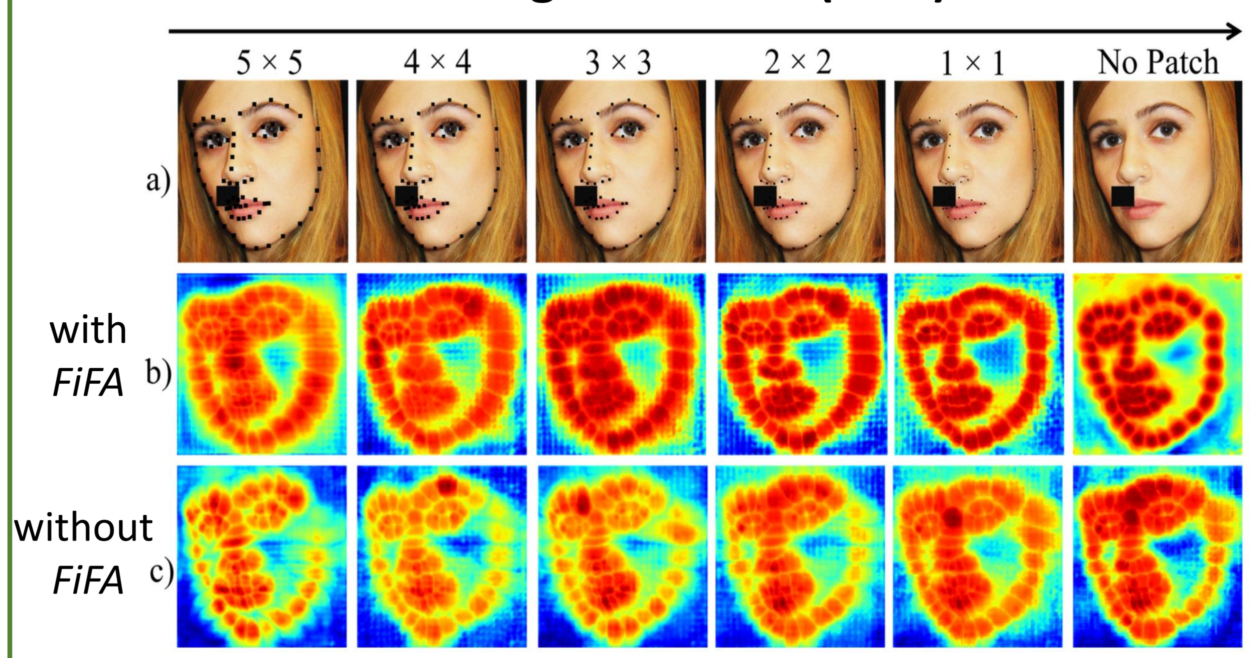

Fig. 2: Illustration of the proposed FiFA.

2. Siamese training with DCCA loss [1]

- To maximize the correlation between two different augmented views which can be expressed as

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

- The DCCA loss ($\mathcal{L}_{DCCA}$) is then computed as

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

3. Transformer + CNN based backbone

- To introduce desirable properties of CNN while retaining characteristics of transformer.
- Anti-aliased CNN [2]: To mitigate the loss of structural information caused by pooling layers.
- FF-Parser layer [3]: To reduce high-frequency noise produced by hour-glass module.

5. Experimental Details:

- Trained/tested on the various benchmark datasets, i.e., WFLW, 300W, COFW and AFLW.
- Along with the DCCA loss, we employ the standard binary cross entropy (BCE) loss and mean absolute error loss for heatmap and coordinate regression, respectively.
- 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_{ic}^{10}), and Area Under the Curve (AUC_{box}).

References:

[1] Galen Andrew et. al., Deep canonical correlation analysis. In International conference on machine learning, pages 1247–1255. PMLR, 2013.

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

[3] Junde Wu et. al., Medsegdiff: Medical image segmentation with diffusion probabilistic model. arXiv preprint arXiv:2211.00611, 2022

[4] Yinglin Zheng et. al., 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.

6. Result Analysis:

Table 1: Effect of method's components on COFW.

Method	NME_{ic} ↓
Vanilla backbone (ViT-B/16) [4]	3.11
+ anti-aliased CNN-based hourglass	3.07
+ Fiducial Focus Augmentation	3.00
+ Siamese training (w DCCA)	2.96

Table 2: Effect of proposed FiFA augmentation technique and Siamese-based DCCA loss on COFW.

Method	Remarks	Baseline	+FiFA	+FiFA + Siamese training (w. DCCA)
HRNET	$ICCV_{21}$	3.45	3.32	**3.11**
ADNet	$ICCV_{21}$	4.68	4.51	**4.45**
FaRL	$CVPR_{22}$	3.11	3.04	**3.01**
SLPT	$CVPR_{22}$	3.32	3.15	**3.10**

Table 3: Comparison with state-of-the-art methods on Pascal COFW, 300W and AFLW datasets.

Method	Remarks	COFW		300W			AFLW			
		NME_{ic} ↓	FR_{ic}^{10} ↓	NME_{ic} ↓			NME_{diag} ↓		NME_{box} ↓	AUC_{box} ↑
				Full	Common	Challenge	Full	Frontal	Full	Full
FaRL	$CVPR_{22}$	3.11	0.12	2.93	2.56	4.45	0.94	0.82	1.33	81.3
SH-FAN	$BMVC_{21}$	3.02	**0.00**	2.94	2.61	4.13	1.31	1.12	2.14	70.0
PropNet	$CVPR_{20}$	3.71	0.20	2.93	2.67	**3.99**	—	—	—	—
SLPT	$CVPR_{22}$	3.32	0.59	3.17	2.75	4.90	—	—	—	—
DTLD	$CVPR_{22}$	3.02	—	2.96	2.60	4.48	1.37	—	—	—
PicassoNet	$TNNLS_{22}$	—	—	3.58	3.03	5.81	1.59	1.30	—	—
FiFA (Ours)	$BMVC_{23}$	**2.96**	**0.00**	**2.89**	**2.51**	4.47	**0.92**	**0.80**	**1.31**	**81.8**

Fig. 3: Qualitative comparison on WFLW benchmark testset. Landmarks shown in green are produced by our method, while the ones in red by the state-of-the-art (SOTA) approach of [4].