Stacked Dense U-Nets with Dual Transformers for Robust Face Alignment

Jia Guo*[1]
https://github.com/deepinsight/insightface

Jiankang Deng*[2]
https://jiankangdeng.github.io/

Niannan Xue[2]
https://ibug.doc.ic.ac.uk/people/nxue

Stefanos Zafeiriou[2]
https://wp.doc.ic.ac.uk/szafeiri/

[1] InsightFace
Shanghai, China

[2] IBUG
Imperial College London
London, UK

Abstract

Facial landmark localisation in images captured in-the-wild is an important and challenging problem. The current state-of-the-art revolves around certain kinds of Deep Convolutional Neural Networks (DCNNs) such as stacked U-Nets and Hourglass networks. In this work, we innovatively propose stacked dense U-Nets for this task. We design a novel scale aggregation network topology structure and a channel aggregation building block to improve the model's capacity without sacrificing the computational complexity and model size. With the assistance of deformable convolutions inside the stacked dense U-Nets and coherent loss for outside data transformation, our model obtains the ability to be spatially invariant to arbitrary input face images. Extensive experiments on many in-the-wild datasets, validate the robustness of the proposed method under extreme poses, exaggerated expressions and heavy occlusions. Finally, we show that accurate 3D face alignment can assist pose-invariant face recognition where we achieve a new state-of-the-art accuracy on CFP-FP (98.514%).

1 Introduction

Facial landmark localisation [7, 8, 24, 25, 28, 34, 35, 37, 38] in unconstrained recording conditions has recently received considerable attention due to wide applications such as human-computer interaction, video surveillance and entertainment. 2D and 3D [1] in-the-wild face alignments are very challenging as facial appearance can change dramatically due to extreme poses, exaggerated expressions and heavy occlusions.

The current state-of-the-art 2D face alignment benchmarks [28, 38] revolve around applying fully-convolutional neural networks to predict a set of landmark heatmaps, where for a given heatmap, the network predicts the probability of a landmark's presence at each

* denotes equal contribution to this work.

[1] In this paper, the 3D facial landmarks refer to the 2D projections of the real-world 3D landmarks, which can preserve face structure and semantic consistency across extreme pose variations.

and every pixel. Since the heatmap prediction for face alignment is essentially a dense regression problem, (1) rich features representations that span resolutions from low to high, and (2) skip connections that preserve spatial information at each resolution, are extensively investigated to combine multi-scale representations to improve inference of where and what [23, 26, 27, 31]. In fact, the most recent state-of-the-art performance in 2D face alignment has been held for a while [35, 38] and is also believed to be saturated [1, 2] by the stacked Hourglass models [26], which repeat resolution-preserved bottom-up and top-down processing in conjunction with intermediate supervision.

Although lateral connections can consolidate multi-scale feature representations in Hourglass, these connections are shallow themselves due to simple one-step operations. Deep layer aggregation (DLA) [36] augments shallow lateral connections with deeper aggregations to better fuse information across layers. We further add the down-sampling paths for the aggregation nodes in DLA and create a new Scale Aggregation Topology (SAT) for network design. Following the same insight in the network topology structure, we propose a Channel Aggregation Block (CAB). The decreasing channel in CAB helps to increase contextual modelling, which incorporates global landmark relationships and increases robustness when local observation is blurred. By combining SAT and CAB, we create the network structure designated dense U-Net. Nevertheless, the computation complexity and model size of the proposed dense U-Net dramatically increases and there is optimisation difficulty during model training especially when the training data is limited. Therefore, we further simplify the dense U-net by removing one down-sampling step as well as substituting some normal convolutions with deep-wise separable convolutions and direct lateral connections. Finally, the simplified dense U-net maintains similar computational complexity and model size as Hourglass, but significantly improves the model's capacity.

Even though stacked dense U-Nets have a high capacity to predict the facial landmark heatmaps, they are still limited by the lack of ability to be spatially invariant to the input face images. Generally, the capability of modelling geometric transformations comes from deeper network design for transformation-invariant feature learning and extensive data augmentation. For transformation-invariant feature learning, Spatial Transform Networks (STN) [21] is the first work to learn spatial transformation from data by warping the feature map via a global parametric transformation. However, such warping is expensive due to additional calculation on explicit parameter estimation. By contrast, deformable convolution [6] replaces the global parametric transformation and feature warping with a local and dense spatial sampling by additional offsets learning, thus introduces an extremely light-weight spatial transformer. For data augmentation, Honari *et al.* [17] have explored a semi-supervised learning technique for face alignment based on having a model predict equivariant landmarks with respect to transformations applied to the image. Similar idea can be found in [33], where mirror-ability, the ability of a model to produce symmetric results in mirrored images, is explored to improve face alignment. Inspired by these works, we innovatively introduce dual transformers into the stacked dense U-Nets. As illustrated in Fig. 1, inside the network, we employ deformable convolution to enhance transformation-invariant feature learning. Outside the network, we design a coherent loss for arbitrary transformed inputs, enforcing the model's prediction to be consistent with different transformations that are applied to the image. With the joint assistance of deformable convolution and coherent loss, our model obtains the ability to be spatially invariant to the arbitrary input face images.

In conclusion, our major contributions can be summarised as follows:

- We propose a novel scale aggregation network topological structure and a channel aggregation building block to improve the model's capacity without obviously increasing

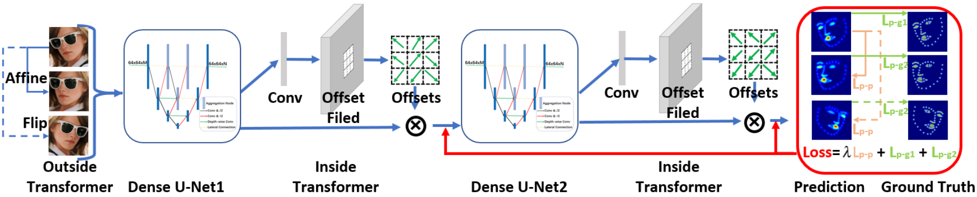


Figure 1: Stacked dense U-Nets with dual transformers for robust facial landmark localisation. We stack two dense U-Nets, each followed by a deformable convolution layer, as the network backbone. The input of the network is one face image together with its affine or flip transformed counterpart. The loss includes heatmap discrepancy between the prediction and ground truth as well as two predictions before and after transformation.

computational complexity and model size.

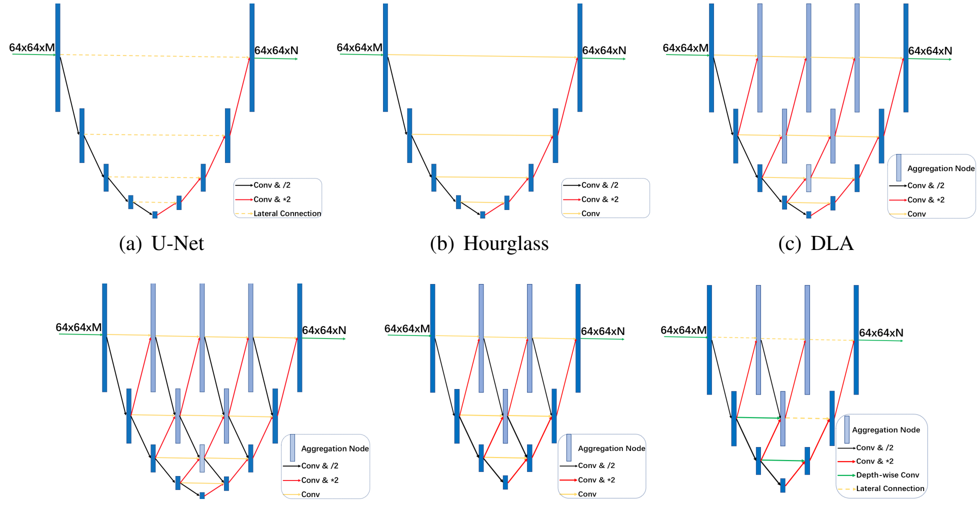
- With the joint assistance of a deformable convolution inside the stacked dense U-Nets and coherent loss for outside data transformation, our model obtains the ability to be spatially invariant to the arbitrary input face images.
- The proposed method creates new state-of-the-art results on five in-the-wild face alignment benchmarks, IBUG [28], COFW [3, 15], 300W-test [28], Menpo2D-test [38] and AFLW2000-3D [40].
- Assisted by the proposed 3D face alignment model, we make a breakthrough in the pose-invariant face recognition with the verification accuracy at 98.514% on CFP-FP [30].

2 Dense U-Net

2.1 Scale Aggregation Topology

The essence of topology design for heatmap regression is to capture local and global features at different scales, while preserving the resolution information simultaneously. As illustrated in Fig. 2(a) and 2(b), the topology of the U-Net [27] and Hourglass [26] are both symmetric with four steps of pooling. At each down-sampling step, the network branches off the high resolution features, which are later combined into the corresponding up-sampling features. By using skip layers, U-Net and Hourglass can easily preserve spatial information at each resolution. Hourglass is similar to U-Net except for the extra convolutional layers within the lateral connections.

To improve the model's capacity, DLA (Fig. 2(c)) iteratively and hierarchically merges the feature hierarchy with additional aggregation nodes within the lateral connections. Inspired by DLA, we further propose a Scale Aggregation Topology (SAT) (Fig. 2(d)) by adding down-sampling inputs for aggregation nodes. The proposed SAT sets up a directed acyclic convolutional graph to aggregate multi-scale features for the pixel-wise heatmap prediction. However, the computation complexity and model size of SAT significantly builds up and the aggregation of three scale signals poses optimisation difficulty during model training especially when the training data is limited. To this end, we remove one step of pooling (Fig. 2(e)), thus the lowest resolution is 8×8 pixels. In addition, we further remove some inner down-sampling aggregation paths and change some normal convolutions into depth-wise separable convolutions [18] and lateral connections [27] as shown in Fig. 2(f). Finally, the



(a) U-Net (b) Hourglass (c) DLA

(d) SAT (I) (e) SAT (II) (f) SAT (III)

Figure 2: Different network topologies. SAT can capture local and global features and preserve spatial information by multi-scale information aggregation.

simplified SAT maintains similar computational complexity and model size as Hourglass, but significantly improves the model's capacity.

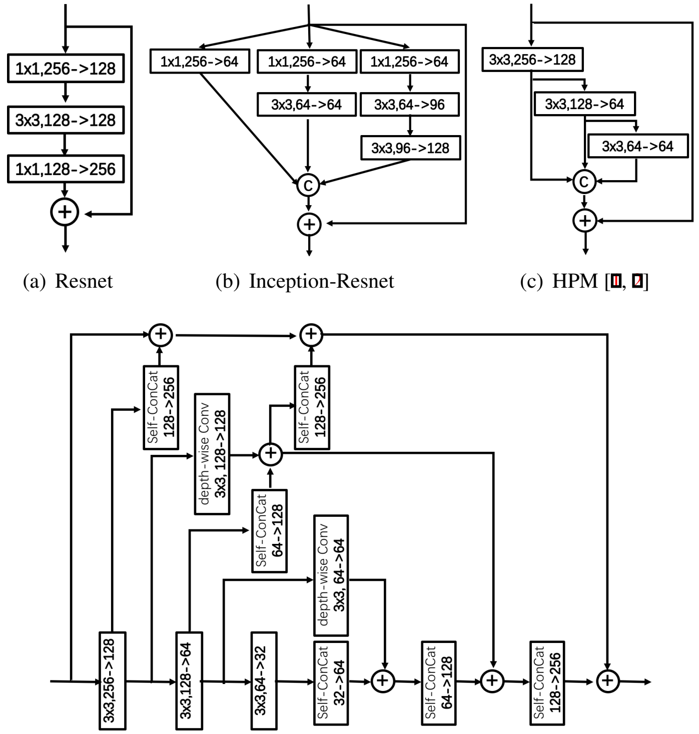
2.2 Channel Aggregation Block

The original Hourglass [26] employs the bottleneck residual block (Fig. 3(a)). To improve the block's capacity, a parallel and multi-scale inception residual block is explored in [12] (Fig. 3(b)). Meanwhile, a novel hierarchical, parallel and multi-scale (HPM) residual block is extensively investigated in [1, 2] (Fig. 3(c)). For the building block design, we follow the same insight in the network topology and innovatively propose a Channel Aggregation Block (CAB). As shown in Fig.3(d), CAB is symmetric in channel while SAT is symmetric in scale. The input signals branch off before each channel decrease and converge back before each channel increase to maintain the channel information. Channel compression in the backbone can help contextual modelling [19], which incorporates channel-wise heatmap relationships and increases robustness when local observation is blurred. To control the computational complexity and compress the model size, depth-wise separable convolutions [18] and replication-based channel extensions are employed within CAB.

3 Dual Transformer

3.1 Inside Transformer

We further improve the model's capacity by stacking two U-Nets end-to-end [2, 26], feeding the output of the first U-Net as input into the second U-Net. Stacked U-Nets with intermediate supervision [26] provide a mechanism for repeated bottom-up, top-down inference allowing for re-evaluation and re-assessment of local heatmap predictions and global spatial configurations. However, stacked U-Nets still lack the transformation modelling capac-

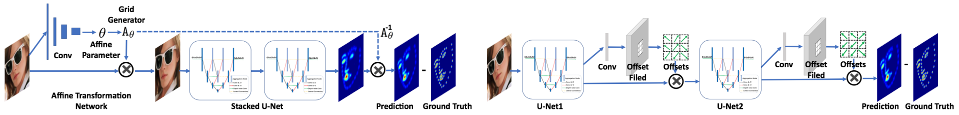


(a) Resnet (b) Inception-Resnet (c) HPM [1, 2]

(d) CAB

Figure 3: Different building blocks. CAB can enhance contextual modelling by channel compression and aggregation.

ity due to the fixed geometric structures. Here, we consider two different kinds of spatial transformers: parameter explicit transformation by STN [21] and parameter implicit transformation by deformable convolution [6]. In Fig. 4(a), we employ the STN to remove the discrepancy of rigid transformation (e.g. translation, scale and rotation) on the input face image, thus the following stacked U-Nets only need to focus on the non-rigid face transformation. Since the variance of the regression target is obviously decreased, the accuracy of face alignment can be easily improved. In Fig. 4(b), the application of deformable convolution behaves the similar way. Nonetheless, the deformable convolution augments the spatial sampling locations by learning additional offsets in a local and dense manner instead of adopting a parameter explicit transformation or warping. In this paper, we employ the deformable convolution as the inside transformer which is not only more flexible to model geometric face transformations but also has higher computation efficiency.

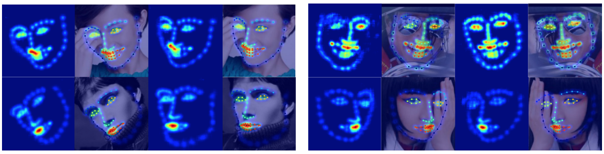


(a) Parameter Explicit Transformation (b) Parameter Implicit Transformation

Figure 4: Inside transformer comparison: STN v.s. Deformable Convolution.

3.2 Outside Transformer

During training, data augmentation by random affine transformation on the input images is widely used to enhance the transformation modelling capacity. Nevertheless, the output heatmaps are not always coherent when there is affine or flip transformation on the input images [17, 33]. As illustrated in Fig. 5(a) and 5(b), there are some obvious local differences between the heatmaps predicted from the original and transformed face images.



(a) Affine Transformation

(b) Flip

Figure 5: Heatmap incoherence under affine and flip transformation applied on the input images.

We explore an outside transformer with an additional loss constraint, which encourages the regression network to output coherent landmarks when there are rotation, scale, translation and flip transformations applied to the images. More specifically, we transform an image during training and enforce the model to produce landmarks that are similarly transformed. Our model is trained end-to-end to minimise the following loss function

$$L = \frac{1}{N}\sum_{n=1}^{N}(\lambda \underbrace{\|H_n(T \odot I) - T \odot H_n(I)\|_2^2}_{L_{p-p}} + \underbrace{\|H_n(I) - G_n(I)\|_2^2}_{L_{p-g1}} + \underbrace{\|H_n(T \odot I) - T \odot G_n(I)\|_2^2}_{L_{p-g2}}), \tag{1}$$

where N is the landmark number, I is the input image, $G(I)$ is the ground truth, $H(I)$ is the predicted heatmaps, T is the affine or flip transformation, and λ is the weight to balance two losses (in Fig. 1): (1) the difference between the prediction and ground truth; and (2) the difference between two predictions before and after transformation.

4 Experiments

4.1 Data

For the training of 2D face alignment, we collate the training sets of the 300W challenge [28] and the Menpo2D challenge [38]. The *300W-train* dataset consists of the LFPW, Helen and AFW datasets. The *Menpo2D-train* dataset consists of 5,658 semi-frontal face images, which are selected from FDDB and ALFW. Hence, a total of 9,360 face images are used to train the 2D68 face alignment model. We extensively test the proposed 2D face alignment method on four image datasets: the *IBUG* dataset (135 images) [28], the *COFW* dataset (507) [3, 15], the *300W-test* dataset (600) [28], and the *Menpo2D-test* dataset (5,535) [38].

For the training of 3D face alignment, we utilise the *300W-LP* dataset [40], which contains 61,225 synthetic face images. The *300W-LP* is generated by profiling and rendering the faces of 300-W [28] into larger poses (ranging from 90° to 90°). We test the proposed 3D face alignment method on the *AFLW2000-3D* dataset (2,000) [40].

Method	IBUG(%)	COFW(%)	Size (mb)	Time (ms)
$Hourglass^{1}$-$Resnet$ [26]	7.32	6.26	13	26
$Hourglass^{2}$-$Resnet$ [26]	7.22	6.18	26	49
$Hourglass^{2}$-$Inception$-$Resnet$	7.07	6.08	38	57
$Hourglass^{2}$-HPM [2]	6.98	5.81	48	47
$Hourglass^{2}$-CAB	6.93	5.77	46	52
$Hourglass^{2}$-HPM (↓ ×3) [2]	6.95	5.82	38	39
$Hourglass^{2}$-CAB (↓ ×3)	6.91	5.78	37	41
U-Net^{2}-CAB	7.17	6.12	36	37
$Hourglass^{2}$-CAB	6.93	5.77	46	52
DLA^{2}-CAB	6.92	5.75	103	61
$SAT(I)\ ^{2}$-CAB	7.05	5.91	131	63
$SAT(II)\ ^{2}$-CAB (↓ ×3)	7.02	5.89	83	47
$SAT(III)\ ^{2}$-CAB (↓ ×3)	6.88	5.74	38	41
$DenseU$-Net + STN	6.81	5.70	116	49
$DenseU$-Net + Inside Transformer	6.77	5.63	49	45
$DenseU$-Net + Outside Transformer	6.80	5.62	38	41
$DenseU$-Net + Dual Transformer	**6.73**	**5.55**	49	45

Table 1: Ablation study for different settings on the IBUG and COFW datasets. Performance is reported as the eye centre distance normalised mean error.

4.2 Training Details

Each face region is cropped and scaled to 128×128 pixels based on the face boxes [39]. We augment the ground truth image with a random combination of horizontal flip, rotation (+/- 40 degrees), and scaling (0.8 - 1.2). The network starts with a 3×3 convolutional layer, followed by a residual module and a round of max pooling to bring the resolution down from 128 to 64, as it could reduce GPU memory usage while preserving alignment accuracy. The network is trained using MXNet with Nadam optimiser, an initial learning rate of 2.5^{-4}, a batch size of 40, and 30k learning steps. We drop the learning rate by a ratio of 0.2 after 16k and 24k iterations. Each training step paralleled on two NVIDIA GTX Titan X (Pascal) takes 1.233s. Although the Mean Squared Error (MSE) pixel-wise loss is given in Eq. 1, in practice we find the Sigmoid Cross-Entropy (CE) pixel-wise loss [1] outperforms the MSE loss for L_{p-g}. Therefore, we employ the CE loss for L_{p-g} and the MSE loss for L_{p-p}, respectively. λ is empirically set as 0.001 to guarantee convergence.

4.3 Ablation Experiments

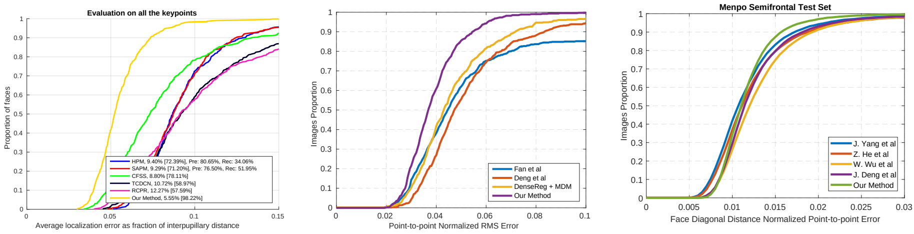
In Tab. 1, we compare the alignment accuracy on the most challenging datasets (IBUG and COFW) under different training settings. We denote each training strategy by $topology^{stack}$-$block$ ($\downarrow \times$ $down$-$sampling\ steps$, 4 by default).

From Tab. 1, we can draw the following conclusions: (1) Compared to a single Hourglass network, a stack of two Hourglass networks can significantly improve the alignment accuracy even though the model size and inference time have been doubled; (2) Building blocks, such as Inception-Resnet, HPM [2] or CAB, can progressively improve the performance with similar model size and computation cost; (3) The performance gap is not apparent between three and four down-sampling steps, but three down-sampling steps can remarkably decrease the model size; (4) As the complexity of the network topology increases from U-

Net, to Hourglass and to DLA, the localisation accuracy gradually raises; (5) Due to limited training data ($\sim 10k$) and the optimisation difficulty with SAT (I) (Fig. 2(d)), e.g. aggregation of three scale signals, the performance obviously drops. By removing some downsampling paths, introducing depth-wise separable convolutions and applying direct lateral convolutions (Fig. 2(f)), the performance bounces back and eventually surpasses Hourglass and DLA; (6) Deformable convolution outperforms STN even with much fewer parameters; (7) Outside transformer with coherent loss can also evidently reduce the alignment error; (8) The proposed stacked dense U-Nets with dual transformers transcend recent state-of-the-art methods on the IBUG (7.02% CVPR18 [14]) and COFW (5.77% CVPR18 [22]) datasets without bells and whistles.

4.4 2D and 3D Face Alignment Results

For 2D face alignment, we further report the Cumulative Error Distribution (CED) curves on three standard benchmarks, that is COFW, 300W-test, Menpo2D-test. On COFW [3] (Fig. 6(a)), our method outperforms the baseline methods in the evaluation toolkit by a prominent margin. The normalised mean error of 5.55% as well as the final success rate at 98.22% are so impressive that the challenge of face alignment under occlusion is even no longer remarkable for our method. In Fig. 7(b), we give some fitting examples on COFW under heavy occlusions. Even by zooming in on these visualised results, we can hardly find any alignment flaw, which indicates that the proposed method can easily capture and consolidate local evidence and global context, and thus improve the model's robustness under occlusions.



(a) COFW (b) 300W-test (c) Menpo2D-test

Figure 6: Landmark localisation results on the COFW, 300W-test, Menpo2D-test datasets. Performance is reported as mean error normalised by the eye centre distance (COFW), the out eye corner distance (300W-test), and the diagonal of the ground truth bounding box (Menpo2D-test), respectively.

On 300W-test [28], we compare our method with leading results, such as Deng *et al.* [7] and Fan *et al.* [13]. Besides, we also compare with the state-of-the-art face alignment method "DenseReg + MDM" [16]. Once again, our model surpasses those methods with ease. On Menpo2D-test [37], we send our alignment results to the organiser and get the CED curves with other best four entries of the competition [38]. As shown in Fig. 6(c), we find our performance is inferior to the best entry [35] within the high accurate interval (NME $< 1.2\%$) because our alignment model is initialised from MTCNN [39] which is less accurate and stable than the detectors applied in [35]. Nevertheless, our model gradually outperforms the best entry, which indicates that our model is more robust under hard cases, such as large pose variations, exaggerated expressions and heavy occlusions.

Method	ESR [5]	RCPR [3]	SDM [32]	3DDFA [40]	HPM [1]	Our Method
NME	7.99	7.80	6.12	4.94	3.26	**3.07**

Table 2: 3D alignment results on the AFLW2000-3D dataset. Performance is reported as the bounding box size normalised mean error [40].

For 3D face alignment, we compare our model on AFLW2000-3D [40] with the most recent state-of-the-art method proposed by Bulat *et al.* [1], which claimed that the problem of face alignment is almost solved with saturated performance. Nonetheless, our method further decreases the NME by 5.8%. In Fig. 7(e), we give some exemplary alignment results, which demonstrate successful 3D face alignment under extreme poses (ranging from $-90°$ to $90°$), accompanied by exaggerated expressions and heavy occlusions.

4.5 3D Face Alignment Improves Face Recognition

Even though face alignment is not claimed to be essential for deep face recognition [9, 10, 11, 29], the face normalisation step is still widely applied in recent state-of-the-art recognition methods [11]. Following the setting of ArcFace [11], we train recognition models under different face alignment methods on VGG2 [4]. As we can see from Tab. 3, the verification performance on LFW [20] is comparatively close. However, on CFP-FP [30], the proposed 3D alignment method obviously decreases the versification error by 48.24% compared to the alignment method proposed in [39]. The significant improvement implies that accurate full-pose face alignment can hugely assist deep face recognition.

Testset	No Alignment	2D-5 landmarks [39]	3D-68 landmarks
LFW	99.63	99.78	**99.80**
CFP-FP	95.428	97.129	**98.514**

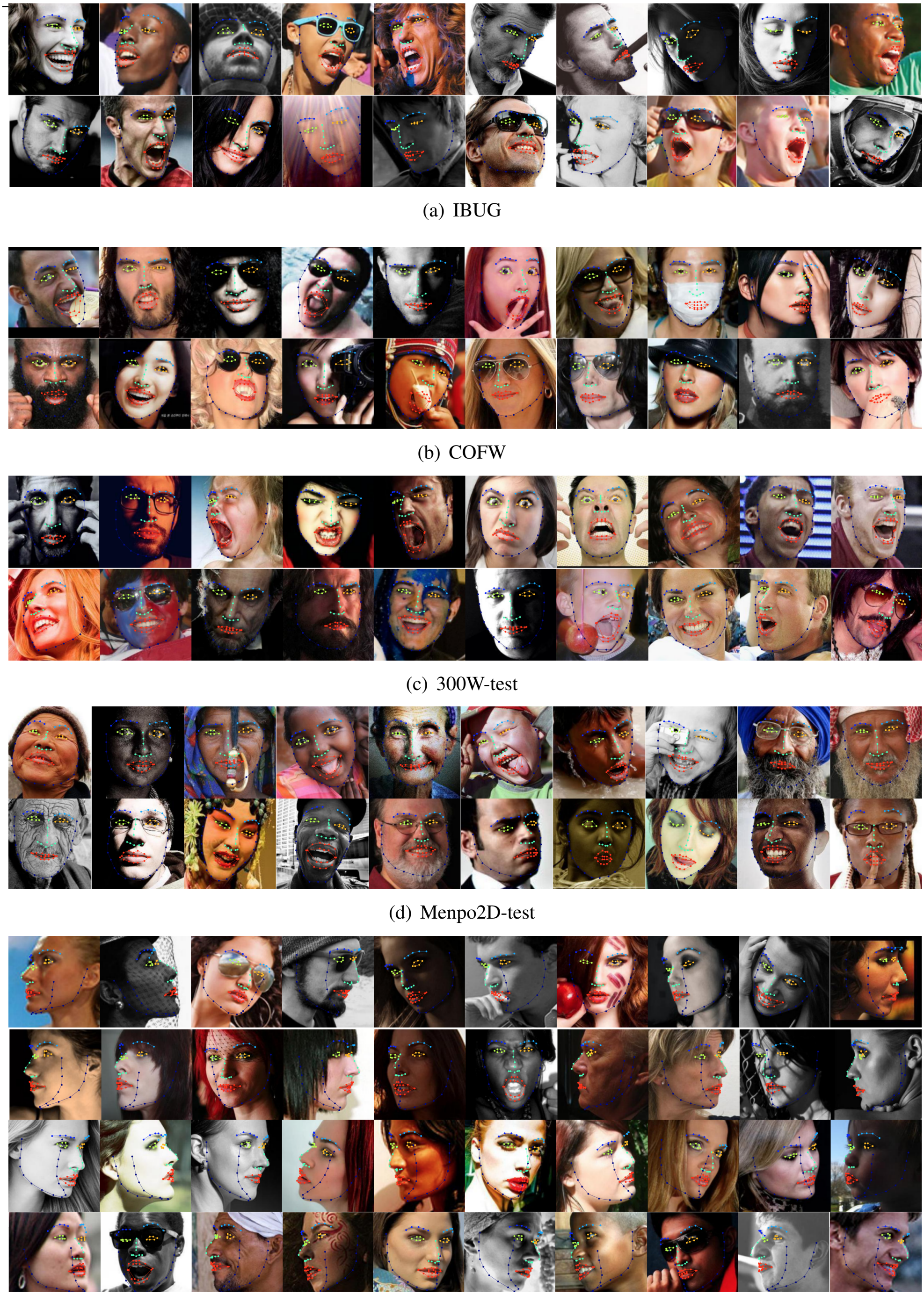
Table 3: Face verification accuracy (%) on the LFW and CFP-FP dataset (ArcFace, LResNet50E-IR@VGG2-LFW-CFP).

5 Conclusion

In this paper, we propose stacked dense U-Nets with dual transformers for robust 2D and 3D facial landmark localisation. We introduce a novel network structure (Scale Aggregation Topology) and a new building block (Chanel Aggregation Block) to improve the model's capacity without sacrificing computational complexity and model size. With the assistance of deformable convolution and coherent loss, our model obtains the ability to be spatially invariant to the input face images. Extensive experiments on five challenging face alignment datasets demonstrate the robustness of the proposed alignment method. The additional face recognition experiment suggests that the proposed 3D face alignment can obviously improve pose-invariant face recognition.

6 Acknowledgement

J. Deng is supported by the President's Scholarship of Imperial College London. This work is also funded by the EPSRC project EP/N007743/1 (FACER2VM) and the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 688520(TeSLA). We thank the NVIDIA Corporation for the GPU donations.



(a) IBUG

(b) COFW

(c) 300W-test

(d) Menpo2D-test

(e) AFLW2000-3D

Figure 7: Example results of 2D and 3D face alignment. The proposed method is robust under pose, expression, occlusion and illumination variations.

References

[1] Adrian Bulat and Georgios Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. *ICCV*, 2017.

[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). *ICCV*, 2017.

[3] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *ICCV*, pages 1513–1520, 2013.

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 2018.

[5] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. In *CVPR*, pages 2887–2894, 2012.

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017.

[7] Jiankang Deng, Qingshan Liu, Jing Yang, and Dacheng Tao. M 3 csr: multi-view, multi-scale and multi-component cascade shape regression. *IVC*, 47:19–26, 2016.

[8] Jiankang Deng, George Trigeorgis, Yuxiang Zhou, and Stefanos Zafeiriou. Joint multiview face alignment in the wild. *arXiv:1708.06023*, 2017.

[9] Jiankang Deng, Yuxiang Zhou, and Stefanos Zafeiriou. Marginal loss for deep face recognition. In *CVPR Workshop*, 2017.

[10] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *CVPR*, 2018.

[11] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv:1801.07698*, 2018.

[12] Jiankang Deng, Yuxiang Zhou, Shiyang Cheng, and Stefanos Zafeiriou. Cascade multiview hourglass model for robust 3d face alignment. In *FG*, 2018.

[13] Haoqiang Fan and Erjin Zhou. Approaching human level facial landmark localization by deep learning. *IVC*, 47:27–35, 2016.

[14] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. *CVPR*, 2018.

[15] Golnaz Ghiasi and Charless C Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. *arXiv:1506.08347*, 2015.

[16] Rıza Alp Güler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. *CVPR*, 2017.

[17] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. *ICCV*, 2017.

[18] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*, 2017.

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *CVPR*, 2018.

[20] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report, 2007.

[21] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NIPS*, pages 2017–2025, 2015.

[22] Amit Kumar and Rama Chellappa. Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. *CVPR*, 2018.

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, volume 1, page 4, 2017.

[24] Qingshan Liu, Jiankang Deng, and Dacheng Tao. Dual sparse constrained cascade regression for robust face alignment. *TIP*, 25(2):700–712, 2016.

[25] Qingshan Liu, Jiankang Deng, Jing Yang, Guangcan Liu, and Dacheng Tao. Adaptive cascade regression model for robust face alignment. *TIP*, 26(2):797–807, 2017.

[26] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

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

[29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[30] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *WACV*, 2016.

[31] Abhinav Shrivastava, Rahul Sukthankar, Jitendra Malik, and Abhinav Gupta. Beyond skip connections: Top-down modulation for object detection. *arXiv:1612.06851*, 2016.

[32] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013.

[33] Heng Yang and Ioannis Patras. Mirror, mirror on the wall, tell me, is the error small? In *CVPR*, 2015.

[34] Jing Yang, Jiankang Deng, Kaihua Zhang, and Qingshan Liu. Facial shape tracking via spatio-temporal cascade shape regression. In *ICCV Workshops*, pages 41–49, 2015.

[35] Jing Yang, Qingshan Liu, and Kaihua Zhang. Stacked hourglass network for robust facial landmark localisation. In *CVPR Workshop*, volume 3, page 6, 2017.

[36] Fisher Yu, Dequan Wang, and Trevor Darrell. Deep layer aggregation. *CVPR*, 2018.

[37] Stefanos Zafeiriou, Grigorios Chrysos, Anastasios Roussos, Evangelos Ververas, Jiankang Deng, and George Trigeorgis. The 3d menpo facial landmark tracking challenge. In *CVPR Workshop*, 2017.

[38] Stefanos Zafeiriou, George Trigeorgis, Grigorios Chrysos, Jiankang Deng, and Jie Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *CVPR Workshop*, 2017.

[39] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *SPL*, 23(10):1499–1503, 2016.

[40] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *CVPR*, pages 146–155, 2016.