Supplementary Material and Related Experiments for BMVC 2022 # 870
CLN: Complementary Learning Network For 3D Face Reconstruction And Dense Alignment

1. Experiments

In this section, folds of experiments are conducted to verify the performance of our CLN. First, we probe into the performance of CLN on the AFLW [5] and AFLW200-3D [18] testing datasets. The training dataset of our network adopts 300W-LP [18]. Then the current 3D face reconstruction and alignment approaches are compared with our CLN. In section 4.3, we conduct ablation studies to determine potential improvements of each module. Finally, we test the generalization of our CLN on LFPW [1] and Helen [6] datasets.

1.1. Datasets

300W-LP [18]: This dataset is an extended version of the 300W [12] across Large Poses. 300W [12] standardises diverse alignment subsets with 68 annotated landmarks, which include AFW [19], IBUG [12], XM2VTS [9], etc. On the basis of 300W [12], 300W-LP [18] is composed of 61225 large pose samples, such as 5207 from AFW and 1786 from IBUG, which is further extended to 122450 images by flipping. Note that the challenging IBUG image set includes large poses, severe occlusion, and exaggerated expression images. In addition to general dataset samples, we also investigate our CLN on challenging samples like IBUG.

AFLW: AFLW [5] means Annotated Facial Landmarks in the Wild. AFLW provides a mass of images collected from Flickr, showing a variety of facial appearances (for example, posture, expression, age, gender, ethnicity) and common imaging and environmental conditions. It contains 25,993 facial images, which are various poses and multiple perspectives in the wild. Each sample has 21 landmarks of annotated information. Due to the complete annotation dataset, AFLW is widely adopted in tasks such as multi-view face detection, face landmark positioning, and face pose estimation. Thus the dataset is well suited for evaluating facial alignment performance in large poses because it includes various large poses yaw from -90° to 90°. Nevertheless, since AFLW only has 21 landmarks annotation, we only use it for 3D landmarks alignment, not for 3D face reconstruction.

AFLW2000-3D: The dataset [18] is used to evaluate challenging facial images in the wild for 3D face alignment. It is composed of the first 2000 facial images of AFLW, and increases the number of labeled landmarks to 68. AFLW2000-3D [18] was first proposed because of the lack of paired 2D images and 3D models in unconstrained environments. Taking into account the recent accomplishments in terms of 3D face reconstruction, 3D face models are constructed from 2D landmarks, and enough 2D landmarks provided can be adopted to accurately fit the 3D model.

Helen [6]: This dataset is a high-resolution face image set collected by Flickr, which contains face images of various poses. It is composed of 2000 training samples and 330 testing facial samples with high-resolution, rich details, and annotations.

LFPW: Compared with Helen [6] dataset, LFPW [1] contains more facial images of expressions, various poses, and different occlusions. LFPW means "Labeled Face Parts in the Wild (LFPW)" gathered from the internet. It contains 1,432 images with 29 landmarks labeled on each image.

1.2. Protocal

Following the previous works, in terms of 3D face alignment and reconstruction, we adopt the normalized mean error to make statistics of the experimental results. The **NME** (Normalized Mean Error) evaluates quantitative results between the geometric shape of the 3D face and the ground-truth face shape to perform quantitative analysis. **CED** curve means statistical analysis of accumulated error results and expressed in quantitative form.

1.3. Experimental details

Our proposed CLN is designed based on Pytorch [10] framework. For each input image of the training dataset, We adopt dual-stream networks to extract features, one of which employs RepVGG to extract holistic information, and the other applies self-shuffling to extract local information. To balance the loss function, our method sets its weight to $\lambda_{\text{wpdc}} = 0.5$ and $\lambda_{\text{wing}} = 1$, respectively. We adopt the SGD [11] optimizer. During the experiment, the batch size and initial learning rate are set to 256 and 0.002, respectively. The input image resolution is 224×224. In our experiments, we train our CLN model by applying 2 NVIDIA RTX 2080Ti GPUs. In parallel mode, GPUs with each available memory of about 11G are used to process the corresponding batch.

1.4. Dense face alignment and reconstruction

For a fair comparison, we estimate the facial alignment effectiveness by employing the normalized mean error (NME) [18] like other methods. The Table 1, 2 reports the NME values for the angles of [0, 30], [30, 60], and [60, 90], namely,

Table 1
The NME(%) of face alignment results on AFLW Dataset (21pts)

Method	[0,30]	[30,60]	[60,90]	Mean	Year
CDM [17]	8.150	13.020	16.170	12.440	2015
RCPR [2]	5.430	6.580	11.530	7.850	2013
ESR [3]	5.660	7.120	11.940	8.240	2014
SDM [15]	4.750	5.550	9.340	6.550	2013
3DDFA [18]	5.000	5.060	6.740	5.600	2016
Yu *et al.* [16]	5.940	6.480	7.960	-	2017
DAMDNet [4]	4.359	5.209	6.028	5.199	2019
GSRN [14]	4.253	5.144	5.816	5.073	2021
MARN [7]	4.306	4.965	5.775	5.015	2021
CLN(Ours)	**3.998**	**4.650**	**5.309**	**4.652**	-

Table 2
The NME(%) of face alignment results on AFLW2000-3D Dataset (68pts)

Method	[0,30]	[30,60]	[60,90]	Mean	Year
RCPR [2]	4.260	5.960	13.180	7.800	2013
ESR [3]	4.600	6.700	12.670	7.990	2014
SDM [15]	3.670	4.940	9.760	6.120	2013
DEFA [8]	4.500	5.560	7.330	5.803	2017
3DDFA [18]	3.780	4.540	7.930	5.420	2016
Yu *et al.* [16]	3.620	6.060	9.560	-	2017
Nonlinear [13]	-	-	-	4.700	2018
DAMDNet [4]	2.907	3.830	4.953	3.897	2019
GSRN [14]	2.842	3.789	4.804	3.912	2021
MARN [7]	2.989	3.670	4.613	3.757	2021
CLN(Ours)	**2.645**	**3.477**	**4.462**	**3.528**	-

small pose, medium pose, and large pose, respectively. The values in bold in each column represent the best performance. The lower the value, the better the result.

Comparison of face alignment on AFLW2000-3D and AFLW: Table 1 tabulates the NME's comparison for the state of the arts approaches and our CLN on AFLW datasets. Figure 1 is the corresponding CED curve. According to the results, The NME value of our CLN achieves 4.652 on the AFLW. Compared with MARN [7], our CLN decrease the error by 0.363 on AFLW. From the results in Table 1, our CLN has a conspicuous boosting in the precision of face alignment for various poses.

We observe in Table 2 that our proposed CLN is superior to other methods. For the AFLW2000-3D dataset (68 points), the three test subsets small pose[0°, 30°], medium pose[30°, 60°], and large pose[60°, 90°] for our CLN algorithm all have an improvement of about 0.335, 0.193, and 0.151 respectively. Finally, the mean NME dropped by 0.229 compared with the MARN [7]. The best results are shown in bold, the lower the value, the better the result. The comparison results between the CED curve results of our proposed CLN and other methods are shown in Figure 1 on AFLW2000-3D. Our CLN is prominently superior to 3DDFA [18], DAMDNet [4], and MARN [7] in the 3D face alignment. In order to compare the alignment results intuitively, we visualize the alignment results as shown in the Figure 3. We show the qualitative comparison results on 3DDFA (second column), DAMDNet (third column), MARN (fourth column), and Our CLN (fifth column). From the last column, our detail results are best at landmarks in the eyebrows, mouth, and cheeks.

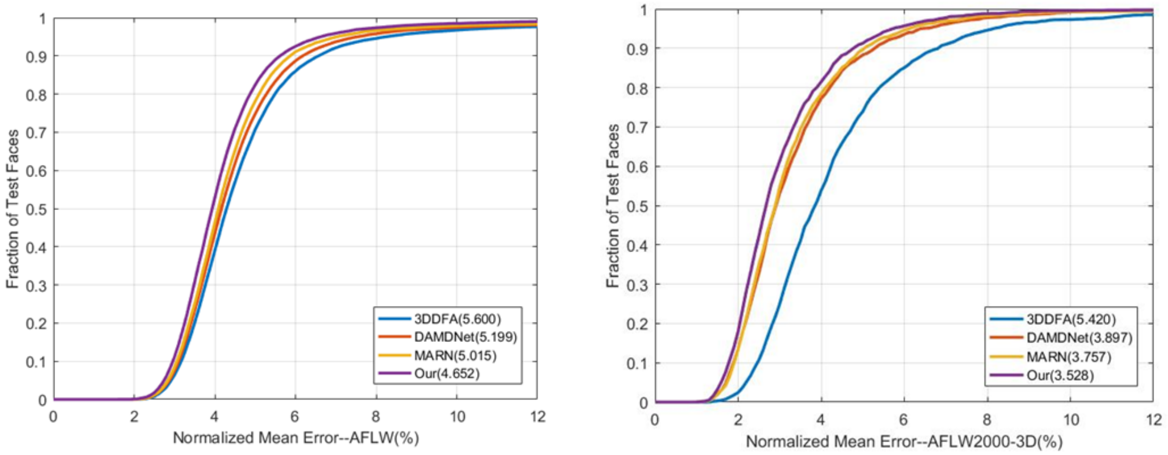

Figure 1: The cumulative errors distribution (CED) curves on AFLW and AFLW2000-3D. Best viewed on a monitor by zooming in.

Table 3
Comparison of our method and other methods on NME(%) for 3D face reconstruction On AFLW2000-3D.

NME_{3D}^{68}	$[0°, 30°]$	$[30°, 60°]$	$[60°, 90°]$	**Mean**
3DDFA	4.877	6.086	8.437	6.467
DAMDNet	4.672	5.619	7.855	6.049
MARN	4.721	5.535	7.483	5.913
CLN(Our)	**4.441**	**5.113**	**7.176**	**5.576**

From the results of Table 1, Table 2, we draw a conclusion that our CLN conquers the problem that the traditional CNN regression methods lack of detailed feature capturing ability in the process of learning large pose faces in the wild. For the input images, Our network not only learns the overall semantic information, but also learns the local semantic information of the local image blocks processed by the self-shuffling module. Through the information integration, so as to achieve a more refined alignment result in the end.

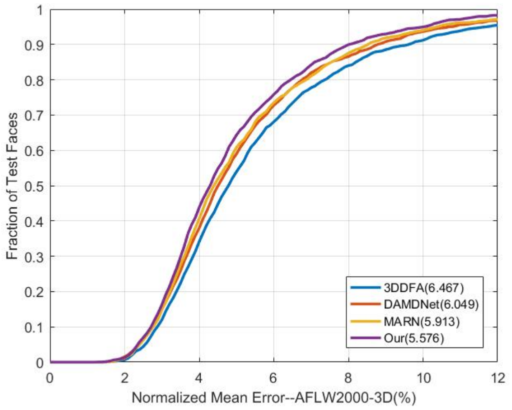

Figure 2: The cumulative errors distribution (CED) curves on AFLW2000-3D.

Comparison of 3D face reconstruction on AFLW2000-3D: Since AFLW only has 21 landmarks annotation, we only use it for 3D landmarks alignment, not for 3D face reconstruction. Thus, we conducted experiments about face reconstruction on AFLW2000-3D. We adopted the 3D NME metric to evaluate the comparison between our CLN and other methods for the 3D face reconstruction on AFLW2000-3D. The comparison is tabulated in Table 3, and the corresponding CED curve is shown in Figure 2. Compared with other approaches, the experimental results prove that our CLN performs better in detail texture

Table 4
Ablation experiments of our CLN.

Model Architecture	*AFLW2000-3D*	*AFLW*
CLN (w/o CDSDA, CAT)	3.763	4.992
CLN (w/o CDSDA)	3.728	4.942
CLN (w/o CAT, No-global)	3.646	4.857
CLN (w/o No-global)	3.624	4.868
CLN (w/o CAT)	3.628	4.845
CLN	**3.528**	**4.652**

and contour. To better demonstrate the effectiveness of our CLN intuitively, Figure 4 shows the comparison of visualization between our CLN and other approaches. We observe that the reconstruction results of our method have more realistic contour details and more natural expressions for large poses and partial occlusions.

For our tasks, face alignment assisted 3D face reconstruction. Since our CLN can capture the local and overall semantic information after self-shuffling processing while extracting the global features of the face, and our CAT modular enhances the network to capture the local similarity facial structure information. Our network alignment results are more accurate, as shown in Figure 3. Therefore, the accurate alignment of the landmarks will lead the model to transform more accurate 3D face model.

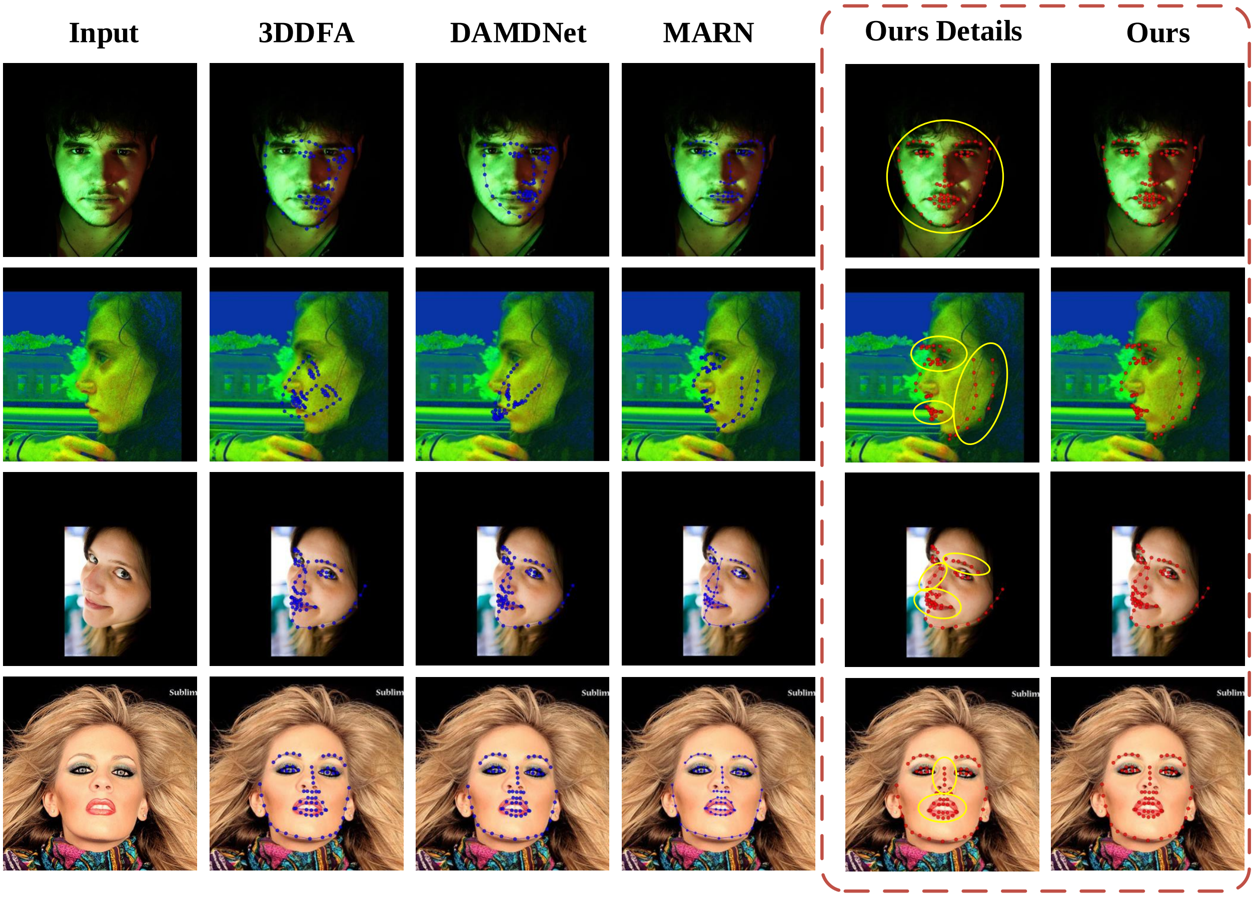

Figure 3: Comparison of 3D face landmarks contours predicted by our network with 3DDFA [18], DAMDNet [4],MARN [7], and CLN(Ours) on AFLW2000-3D, our (CLN) details are shown in the last column.

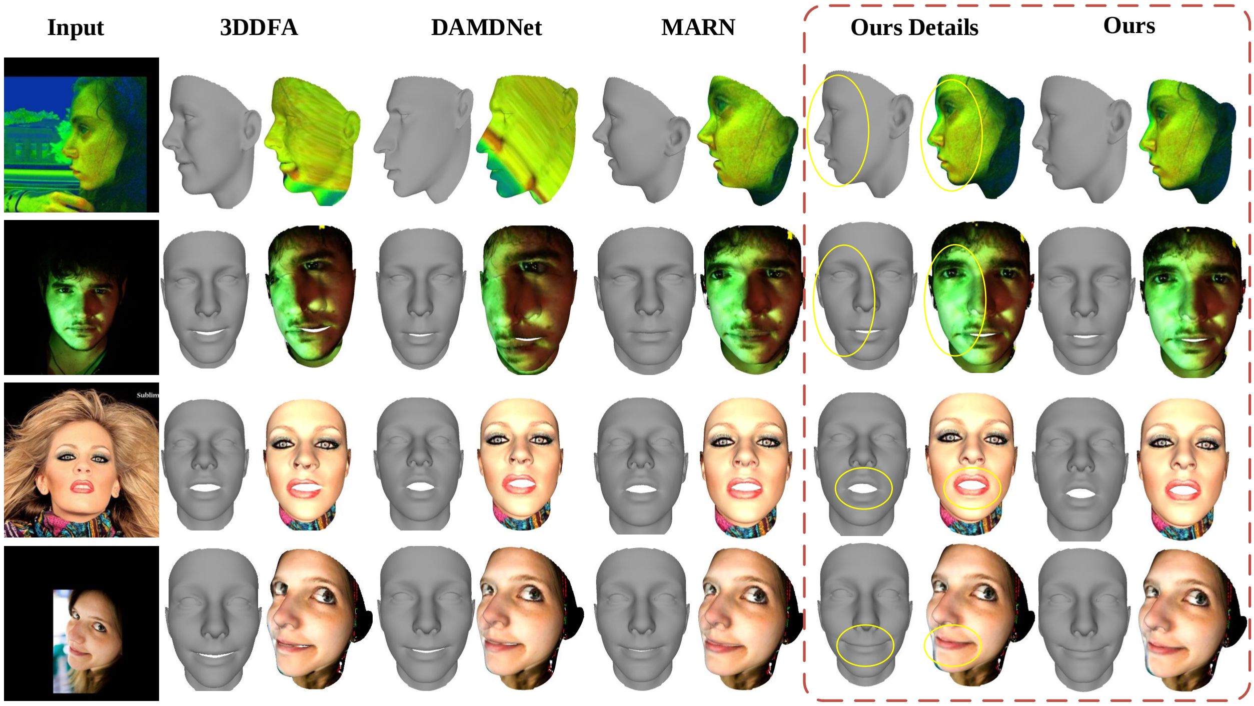

Figure 4: Comparison of 3D face models predicted by our network with 3DDFA [18], DAMDNet [4],MARN [7], and CLN(Ours) on AFLW2000-3D. Our (CLN) details are shown in the last column.

1.5. Ablation study

In this subsection, to evaluate the effectiveness of our proposed CLN, we conducted several ablation experiments. We tested how various components affected the performance of the model, focusing on the effects of the cross-domain self-shuffling data augmentation (CDSDA), shuffled image with only local information (No-global), and coordinate attention transformer (CAT). Under various experimental settings, we adopted NME (%) to measure the effectiveness of the different modules on AFLW2000-3D and AFLW.

Table 4 shows the effectiveness of various components on our CLN model. The CDSDA branch randomly shuffles and reorganizes the original pictures combined with a thumbnail for enhancing the learning ability of the network through data augmentation. The CAT module employs the captured position dependence of the spatial hierarchy and the precise position information of the spatial direction to help the entire model reposition. The global module uses smaller thumbnails to replace empty blocks instead of the mask. Even if some local information is covered, the information will not be lost due to the presence of global thumbnail information. At the same time, it can be better retained contextual information during feature extraction. Global information enhances the network's ability to extract global features.

1.6. Extended Experiment

To verify the robustness and generalization of our CLN in the wild and unconstrained scenes, we provide additional experiments on Helen [6] and LFPW [1] testing datasets with the trained CLN model. We explore experiments on Helen, and the qualitative outcomes are shown in Figure 6 and Figure 5. Figure 6 shows the outcomes of 3D face reconstructed models, our models are superior to 3DDFA, DAMDNet, and MFIRRN in terms of contour and expression. Especially for special face shapes, such as the faces of children in the second and fourth rows, the face models reconstructed by DAMDNet and MFIRRN are too average. Although the overall outline of our model is not much different from 3DDFA, the expression is more realistic, as can be seen from the degree of opening and closing of the mouth. In Figure 5, we observe that our CLN achieves more precise landmarks' location compared with the prior works. In detail, we use yellow ellipses to mark the details of our CLN method for positioning the landmarks. Our method is more accurate in predicting the corners of the mouth, and the landmarks on the cheeks of people with partial occlusion are more realistic.

Following the experiments on Helen, we also conduct analogous experiments on LFPW [1]. The 3D face alignment and reconstructed models results are shown in Figure 5 and Figure 7, respectively. Figure 5 shows our details in yellow ellipses. Compared with other methods, our CLN extract rich details and predict more accurately from landmarks alignment results about eyes, corners of mouth, and chin. As shown in Figure 7, 3D reconstructed visualizations using our CLN have finer textures and contours. Benefitting from partial information reorganization, CLN can recover more weak texture regions.

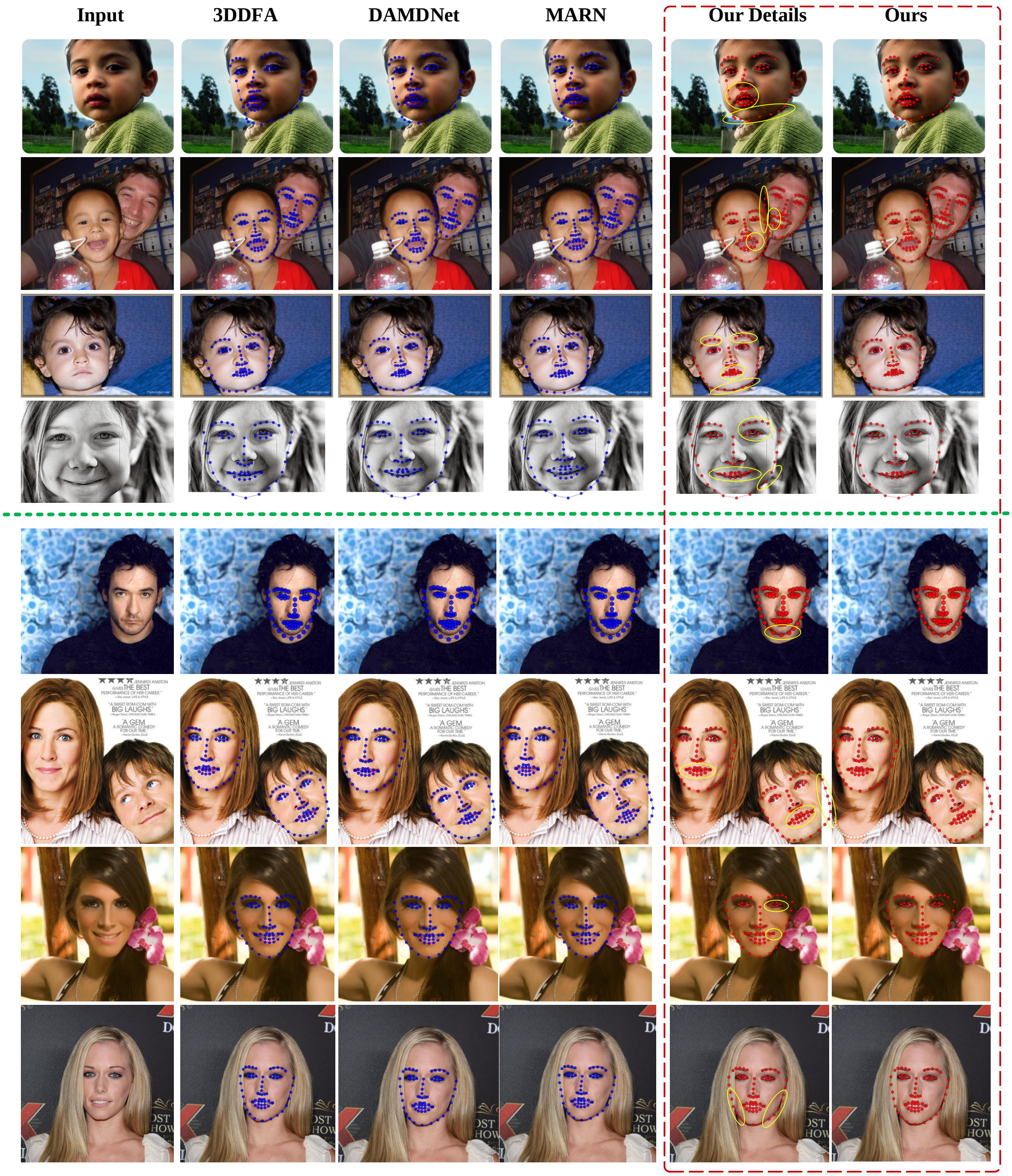

Figure 5: Qualitative results of 3D face alignment on Helen and LFPW. The input image is shown in the first column, the other columns show the results for our method compared to various baselines. The first four rows show the test results on Helen, and the last four rows show the test results on LFPW. Please zoom in to view details.

2. Conclusion

In this paper, we propose a complementary learning network for 3D face reconstruction from unconstrained samples, which highlights the importance of complementary learning between global and local information. Specifically, we elaborately design a cross-domain self-shuffling data augmentation method to embed global information into complex local features and make full use of local features to enrich geometric details. Then, we design a dual-stream network to jointly learn global discriminative features and local detailed features. By doing this, we achieve the mutual complement between the two domain data and utilize the dual-stream fusion network to achieve the mutual complement between the global discriminative features and the local detailed features. Therefore, our network is able to reconstruct a 3D face shape with rich details in

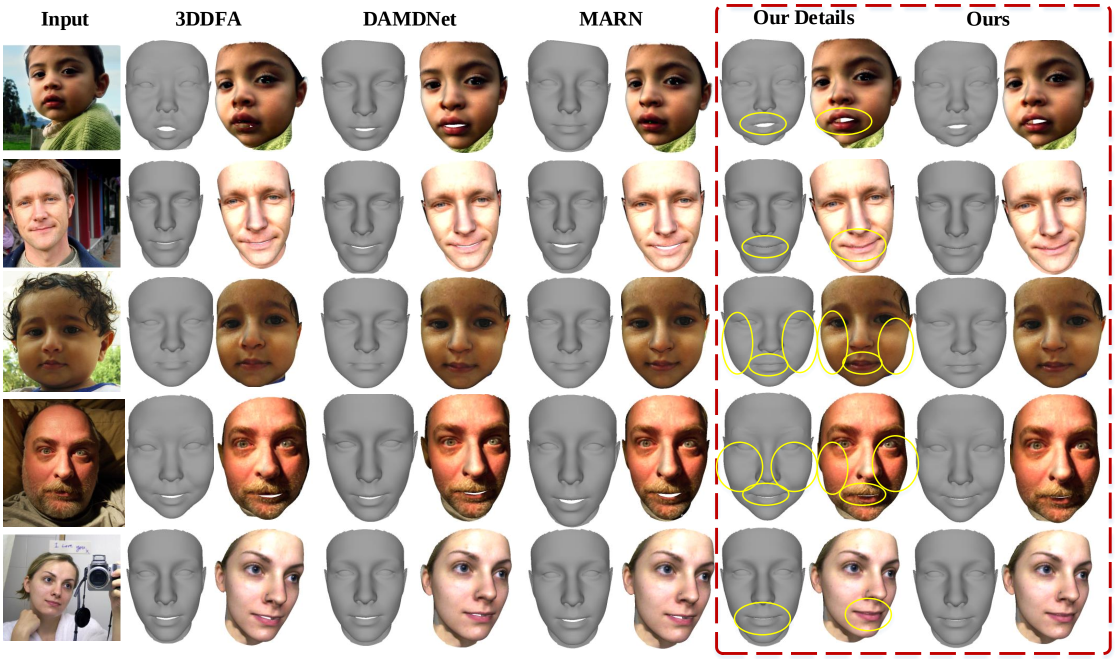

Figure 6: Qualitative results of 3D face reconstruction on Helen. The input image is shown in the first column, the other columns show the results for our method compared to various baselines.

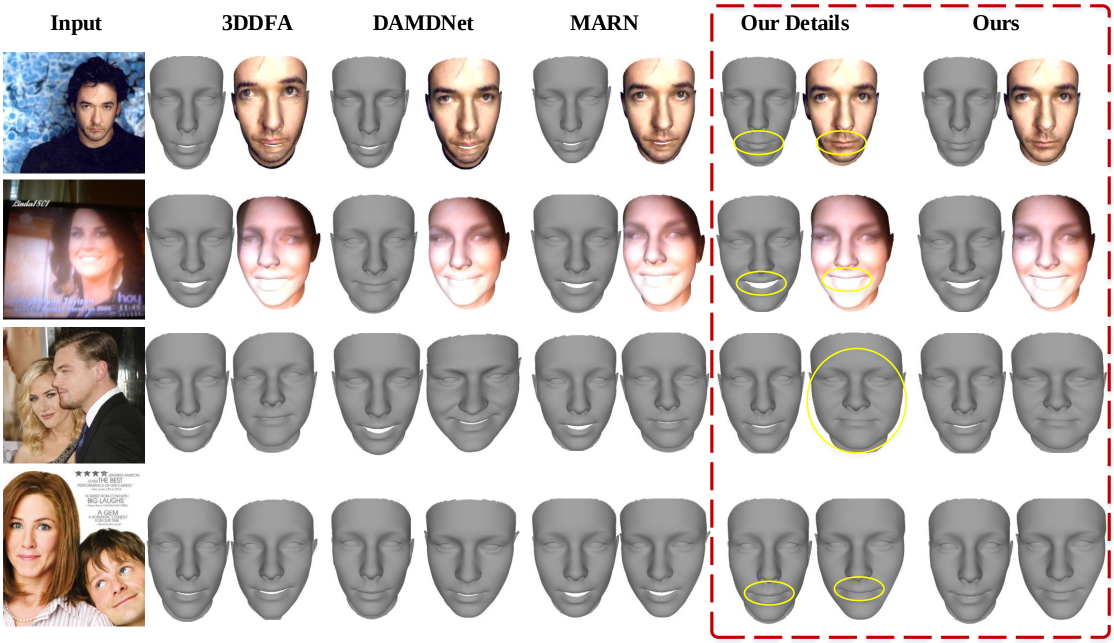

Figure 7: Qualitative results of 3D face reconstruction on LFPW. The input image is shown in the first column, the other columns show the results for our method compared to various baselines. The last two input images have two faces, and the reconstructions are displayed from left to right.

a manner of complementary learning. Moreover, to improve the detection accuracy for fine face reconstruction, we adopt the coordinate attention transformer module to enforce our network's ability of capturing the correlation between local information. From folds of experiments, we conclude that our CLN approach has achieved great boosting in both dense face alignment and 3D face reconstruction. In future work, We further explore the geometry to constraint face reconstruction task by fitting landmarks.

References

[1] Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N., 2013. Localizing parts of faces using a consensus of exemplars. IEEE transactions on pattern analysis and machine intelligence 35, 2930–2940.

[2] Burgos-Artizzu, X.P., Perona, P., Dollár, P., 2013. Robust face landmark estimation under occlusion , 1513–1520.

[3] Cao, X., Wei, Y., Wen, F., Sun, J., 2014. Face alignment by explicit shape regression. International Journal of Computer Vision 107, 177–190.

[4] Jiang, L., Wu, X.J., Kittler, J., 2019. Dual attention mobdensenet (damdnet) for robust 3d face alignment , 0–0.

[5] Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H., 2011. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization , 2144–2151.

[6] Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S., 2012. Interactive facial feature localization, in: European conference on computer vision, Springer. pp. 679–692.

[7] Li, X., Wu, S., 2021. Multi-attribute regression network for face reconstruction , 7226–7233.

[8] Liu, Y., Jourabloo, A., Ren, W., Liu, X., 2017. Dense face alignment , 1619–1628.

[9] Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G., et al., 1999. Xm2vtsdb: The extended m2vts database, in: Second international conference on audio and video-based biometric person authentication, Citeseer. pp. 965–966.

[10] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in pytorch .

[11] Qian, N., 1999. On the momentum term in gradient descent learning algorithms. Neural networks 12, 145–151.

[12] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M., 2013. 300 faces in-the-wild challenge: The first facial landmark localization challenge, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 397–403.

[13] Tran, L., Liu, X., 2018. Nonlinear 3d face morphable model , 7346–7355.

[14] Wang, X., Li, X., Wu, S., 2021. Graph structure reasoning network for face alignment and reconstruction, in: International Conference on Multimedia Modeling, Springer. pp. 493–505.

[15] Yan, J., Lei, Z., Yi, D., Li, S., 2013. Learn to combine multiple hypotheses for accurate face alignment , 392–396.

[16] Yu, R., Saito, S., Li, H., Ceylan, D., Li, H., 2017. Learning dense facial correspondences in unconstrained images , 4723–4732.

[17] Yu, X., Huang, J., Zhang, S., Metaxas, D.N., 2015. Face landmark fitting via optimized part mixtures and cascaded deformable model. IEEE transactions on pattern analysis and machine intelligence 38, 2212–2226.

[18] Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z., 2016. Face alignment across large poses: A 3d solution , 146–155.

[19] Zhu, X., Ramanan, D., 2012. Face detection, pose estimation, and landmark localization in the wild, in: 2012 IEEE conference on computer vision and pattern recognition, IEEE. pp. 2879–2886.