FAB: A Robust Facial Landmark Detection Framework for Motion-Blurred Videos

Supplementary Materials

Keqiang Sun[1]*, Wayne Wu[2], Tinghao Liu[3], Shuo Yang[4]
Quan Wang[3], Qiang Zhou[2], Zuochang Ye[1], Chen Qian[3]
[1]Institute of Microelectronics, Tsinghua University
[2]Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University
[3]SenseTime Research [4]Amazon Rekognition
{skq17,wwy15}@mails.tsinghua.edu.cn, {zuochang,zhouqiang}@tsinghua.edu.cn,
{liutinghao,wangquan,qianchen}@sensetime.com, shuoy@amazon.com

Abstract

This supplementary materials is mainly composed of four parts. Firstly, we provided the detailed architectures of our proposed algorithm in the main paper, in which the training details are presented as well. Secondly, we conducted a comprehensive evaluation of state-of-the-art algorithms on Blurred-300VW, and RWMB datasets. Thirdly, we further introduced the proposed benchmark for video with real-world motion blur (RWMB), and illustrated the superior and necessity of this dataset. Lastly, in the appendix, we make a video to evaluate our algorithm on realistic videos. All the datasets (Blurred-300VW, RWMB), models and codes of this work will be released.

1. Architecture and training

The whole framework of our proposed algorithm is shown and introduced in the main paper. In this section, we present the detailed architecture of our framework for better understanding.

1.1. Structure predictor

Architecture The main function and working flow of the Structure Predictor is shown in Figure 1.

The main architecture of hourglass employed in Structure Predictor is introduced in Figure 2. The hourglass has mirrored encoding (convolution and max-pooling) and decoding (convolution and up-sampling) architectures. Skip-connections shown in the Figure 2 provide connections between the decoding and encoding feature maps with the same size. The corresponding feature maps are fused using element-wise addition. The inputs of the Hourglass are normalized to (0,1) using the max pixel value.

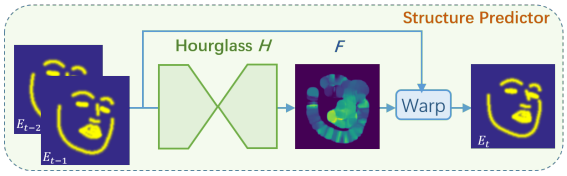

Figure 1: Structure Predictor. The Structure Predictor takes two previous face boundaries as inputs. An hourglass is used to form the optical flow, and the warping block warp the previous two boundaries into a new one, which will aid the motion deblurring in the Structure-aware Motion Deblurring module

Layer Type	Kernel Size	Ourput Channels	Output size
Input	n/a	2	256
Convolution	3	64	256
Pooling	2	128	128
Convolution	3	128	128
Pooling	2	256	64
Convolution	3	256	64
Pooling	2	256	32
Convolution	3	256	32
Upsampling	2	256	64
Convolution	3	256	64
Upsampling	2	128	128
Convolution	3	128	128
Upsampling	2	64	256
Convolution	3	64	256
Convolution	3	2	256

Figure 2: Hourglass employed in our architecture.

*The work was done during the internship at SenseTime Research.

The warping block is an interpolation function [5] to generate next face edge. Given previous two face edges E, the output of the warping block is:

$$\boldsymbol{W}(E,F) = \sum_{i,j,k\epsilon[0,1]} W^{ijk} E(V^{ijk}) \tag{1}$$

where V are eight vertices of the pixel in the input frames:

$$\begin{aligned} V^{000} &= (\lfloor L_x^0 \rfloor, \lfloor L_y^0 \rfloor, 0) \\ V^{100} &= (\lceil L_x^0 \rceil, \lfloor L_y^0 \rfloor, 0) \\ &\dots \\ V^{011} &= (\lfloor L_x^1 \rfloor, \lceil L_y^1 \rceil, 1) \\ V^{111} &= (\lceil L_x^1 \rceil, \lceil L_y^1 \rceil, 1) \end{aligned} \tag{2}$$

where L^0 and L^1 are defined as the absolute coordinates of the corresponding location of the first and second input frames. And W is the trilinear re-sampling weight:

$$\begin{aligned} W^{000} &= (1-(L_x^0 - \lfloor L_x^0 \rfloor))(1-(L_y^0 - \lfloor L_y^0 \rfloor)) \cdot \frac{1}{3} \\ W^{100} &= (L_x^0 - \lfloor L_x^0 \rfloor)(1-(L_y^0 - \lfloor L_y^0 \rfloor)) \cdot \frac{1}{3} \\ &\dots \\ W^{011} &= (1-(L_x^1 - \lfloor L_x^1 \rfloor))(L_y^1 - \lfloor L_y^1 \rfloor) \cdot \frac{2}{3} \\ W^{111} &= (L_x^1 - \lfloor L_x^1 \rfloor)(L_y^1 - \lfloor L_y^1 \rfloor) \cdot \frac{2}{3} \end{aligned} \tag{3}$$

where $\frac{1}{3}$ and $\frac{2}{3}$ corresponds to the weight of the first and second input frame in the predicted frame.

Training For Structural Predictor pretraining, facial edges are implemented as inputs. Set current time as t, ground truth landmarks L_{t-1} at last time point $t-1$ are converted to 256×256 image E_{t-1} following the method proposed in [look at boundary], and the original E_{t-1} is pushed one step forward as E_{t-2}. E_{t-1} and E_{t-2} are concatenated and feed to Structural Predictor to predict E_t. We fit annotated landmarks face edges by cubic spline interpolation and use them as input and ground truth. Gaussian distribution with standard deviation of 0.01 is used to initialize the weights. The basic learning rate is 0.001 and decays to its 0.96 every $2,000$ steps. The Structural Predictor is trained $500,000$ steps totally.

1.2. Structure-aware motion deblurring module

Architecture The main function and working flow of the Structure-aware motion deblurring module is shown in Figure 3.

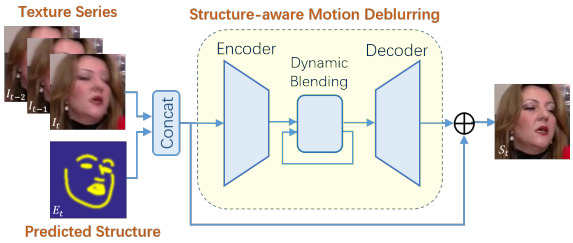

Figure 3: Structure-aware Motion Deblurring Module. A batch of three recent frames and predicted boundary map concatenate as input. An encoder network extracts needed information from inputs. Then we use a dynamic temporal blending network to combine information across different frames. Finally, a decoder is used to predicts the residual between blurry frames and ground truth.

Layer Type	Kernel Size	Ourput Channels	Output size
Input	-	7	256
Convolution	5	64	256
Convolution	3	32	128
Concatenate	-	-	128
Residual	3	64	128
Residual	3	64	128
Residual	3	64	128
Residual	3	64	128
Concatenate	-	-	128
Convolution	3	64	128
Weighted Mean	-	-	128
Residual	3	64	128
Residual	3	64	128
Residual	3	64	128
Residual	3	64	128
Deconv	4	64	128
Convolution	3	3	256
Sum	-	3	256

Figure 4: Architecture of the Structure-aware Motion Deblurring network.

The detailed architecture of the Structure-aware Motion Deblurring Module is illustrated in the Figure 4. The Deblurring Module is composed of an Encoder, a Decoder and a Dynamic Blending module. In the Encoder module, four Residual blocks are utilized to extract information from inputs. And in the Decoder module, also four Residual blocks are utilized to combine information across frames.

As is shown in Figure 5, residual block mentioned in the Structure-aware Motion Deblurring Module consists two convolution layer. Skip connection add the input and the input and the output of convolution layer element-wise.

Training For Structure-aware Motion Deblurring pretraining, annotation results for current frame is interpolated to face edge and used as structural information. Face edges

Layer Type	Kernel Size	Ourput Channels	Output size
Input	-	C	S
Convolution	K	C	S
BN	-	-	S
Convolution	K	C	S
BN	-	-	S
ReLU	-	-	S

Figure 5: Hourglass employed in our architecture.

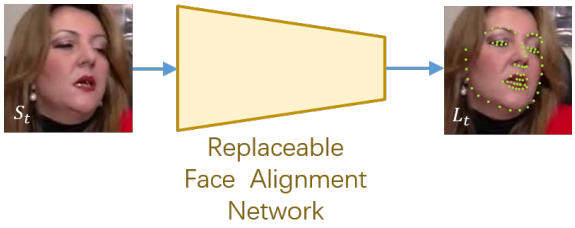

Figure 6: Pre-activation Resnet-18 plays the role of the facial landmark detector in our framework. By using the deblurred sharp image as input, the Pre-activation Resnet-18 predicts the landmark location.

and picture series are concatenated and serve as input. Basic learning rate is 0.0001 and decays to its 0.96 every $2,000$ steps. The Structural Predictor is trained $500,000$ steps totally.

1.3. Pre-activation Resnet

Architecture In our main paper, we take pre-activation Resnet-18 [3, 4] as the facial landmark detector. The architecture if the pre-activation Resnet used in our paper is shown in Figure 6. The layer details of the pre-activation Resnet is presented in Figure 7.

As is shown in Figure 7, the pre-activation Resnet-18 is mainly composed of four Pre-act Res-Block, which is further explained in Figure 8.

By putting the ReLU and BN layer forward, the optimization is further eased and using BN as pre-activation improves regularization of the models. These would benefit the training process and improve the generalization ability.

Training For facial landmark detector pretraining, the network weights is initialized by MSRA weight initialization method. Firstly, we trained the model on 300W-dataset for $1,000,000$ steps, basic learning rate is 0.0002, and decays to its 0.5 every $200,000$ steps, after that,we fine-tuned the network on the 300VW dataset and 300W-dataset for $500,000$ steps, reduce the base learning rate to 0.00002, and also decays to its 0.5 every $100,000$ steps. Of course, some data augmentation such as trans, rotate, zoom and mirror is also used in the total training stage.

Layer Type	Kernel Size	Output Channels	Output size
Input	-	3	256
Convolution	7	16	128
BN	-	16	128
ReLU	-	16	128
Max Pooling	3	16	64
BN	-	16	128
ReLU	-	16	128
Pre-act ResBlock	-	8	128
Pre-act ResBlock	-	16	64
Pre-act ResBlock	-	32	32
Pre-act ResBlock	-	64	16
FC+ReLU	-	256	-
FC+ReLU	-	256	-
FC	-	136	-

Figure 7: Overall architecture of the Pre-activation Resnet-18.

Layer Type	Kernel Size	Output Channels	Output size
Input	-	C	S
Convolution	1	C	S
BN	-	C	S
ReLU	-	C	S
Convolution	3	C	S
BN	-	C	S
ReLU	-	C	S
SUM	-	C	S
BN	-	C	S
ReLU	-	C	S
Convolution	-	C	S
BN	-	C	S
ReLU	-	C	S
Convolution	-	C	S
BN	-	C	S
ReLU	-	C	S
Convolution	-	C	S
SUM	-	C	S

Figure 8: According to the [4], the "full pre-activation", namely putting both ReLU and BN before the convolution layer in the residual block, is a better deformation of the original Residual Network.

2. Evaluation of state-of-the-art algorithms

In this section, we conducted the first, to the best of our knowledge, comprehensive evaluation of state-of-the-art and classical algorithms on 300VW, Blurred-300VW and RWMB datasets. We retrained these algorithms with identical datasets for fair comparison. We used training strategies and hyper-parameters introduced by the authors.

Method	NME		
	Category A	Category B	Category C
FAN [1]	10.8498	6.8068	6.1044
RFLD [6]	10.9602	8.1978	4.6273
SAN [2]	5.6192	7.7097	4.3380
SDM [10]	10.4043	8.5230	4.8287
TCDCN [11]	6.8172	7.8907	4.3506
LAB [9]	5.2837	6.0719	3.9563
Ours	**4.2396**	**5.6657**	**3.1624**

Table 1: Mean error normalized by inter-ocular distance, on Blurred-300VW dataset

Method	NME		
	Category A	Category B	Category C
FAN [1]	15.6887	9.6413	8.6445
RFLD [6]	15.8117	11.6122	6.5511
SAN [2]	8.1786	10.9207	6.1393
SDM [10]	14.6990	12.0512	6.8066
TCDCN [11]	9.7781	11.1749	6.1578
LAB [9]	7.7321	8.6009	5.5996
Ours	**6.1842**	**8.0253**	**4.4750**

Table 2: Mean error normalized by inter-pupil distance, on Blurred-300VW dataset

Method	NME		
	Category A	Category B	Category C
FAN [1]	4.0833	2.8543	2.3645
RFLD [6]	4.1517	3.4501	1.7822
SAN [2]	2.0466	3.2280	1.6691
SDM [10]	3.4984	3.5356	1.8914
TCDCN [11]	2.5020	3.2999	1.6734
LAB [9]	1.9287	2.5439	1.5304
Ours	**1.5314**	**2.3755**	**1.2208**

Table 3: Mean error normalized by the diagonal length of the bounding box, on Blurred-300VW dataset

Method	NME		
	inter-ocular	inter-pupil	diagonal length
FAN [1]	13.7418	19.1374	5.1479
RFLD [6]	16.3358	22.7582	6.0383
TCDCN [11]	10.4791	14.6000	3.8468
SAN [2]	10.3991	14.4885	3.8397
LAB [9]	9.4717	13.1945	3.5258
Ours	**8.4317**	**11.7455**	**3.1445**

Table 4: Mean error normalized by the inter-ocular distance, inter-pupil distance and diagonal length of the bounding box, on RWMB dataset

In previous works, differentiated normalization methods are leveraged in different papers [9, 1, 7]. Disunited normalization methods have make it inconvenient for following researchers to compare results. Therefore, we reimplemented these state-of-the-art and classical algorithms [10, 11, 1, 6, 2, 9], and reported Normalized Mean Error(NME) normalized with inter-ocular distance, inter-pupil distance and diagonal length of the bounding box. We also reported Failure Rate and Area Under Curve(AUC) with several NME thresholds. Corresponding Cumulative Error Distribution (CED) curves are reported in this supplementary material as well to set complete baselines for future works.

2.1. Evaluation on Blurred-300VW

In this section, we present our evaluation results on Blurred-300VW dataset.

Blurred-300VW is a dataset we generated from original 300VW, following the blurring method introduced in [8]. 20 subframes are generated according to the optical flow. Then mean value of these 20 subframes is calculated to mimic motion blur. Annotation of each generated frame are taken from the middle-time subframe. We show some blurred pictures in Figure 9.

NME value is normalized with inter-ocular distance, inter-pupil distance and diagonal length of the bounding box. Failure Rate, AUC with various thresholds, and CED curves are presented.

As is shown in Figure 10, 11 and 12, our algorithm performs state-of-the-art in Category A and Category C with a great margin. Since the Blurred-300VW data set contains more blurry images, our algorithm outperforms others with even greater margin in all three categories. NME is shown in Table 1, 2 and 3, Failure Rate and AUC are shown in Table 5, 6, 7, 8, 9, 10, 11, 12 and 13. In each table, mean error is normalized with different methods. Also, AUC and failure rate are counted with various threshold values.

We also compare different deblurring and facial landmark detection effect in Figure 15 and 14. Images in the first column are samples of Blurred-300VW. Landmarks of these blurry faces are presented in the second column. Red points are ground truth facial landmarks of the input and the green points are outputs of state-of-the-art facial landmark detection algorithms. By naively applying state-of-the-art deblurring and facial landmark detection algorithm, we got column three and column four. "Naive" here means a direct and simple way of concatenation. Lastly, the proposed algorithm in the main paper produces the deblurred picture in column five and landmarks in column six.

Figure 9: Blurred-300VW generated by interpolating frames and take the mean value. The first and third rows are original 300VW, the second and forth rows are generated Blurred-300VW.

2.2. Evaluation on RWMB

In this section, we present our evaluation on RWMB dataset.

NME value is normalized with inter-ocular distance, inter-pupil distance and diagonal length of the bounding box. Failure Rate, AUC with various thresholds and CED curves are presented.

As is shown in Figure 13, our algorithm performs state-of-the-art in our proposed dataset. We also schedule to release a training set containing more videos with real-world motion blur in order to make this blurry dataset self-containing. NME is shown in Table 4, Failure Rate and AUC are shown in Table 14, 15 and 16. In each table, mean error is normalized with different methods. Also, AUC and failure rate are counted with various threshold values.

Method	Category A		Category B		Category C	
	Failure Rate(%)	AUC	Failure Rate(%)	AUC	Failure Rate(%)	AUC
FAN [1]	3.7456	0.6107	0.8876	0.6823	**0.0533**	0.6964
RFLD [6]	14.1309	0.6878	7.4951	0.6827	1.6516	0.7875
SAN [2]	0.8038	0.7438	**0.0000**	0.6145	0.6390	0.7905
TCDCN [11]	1.4394	0.6815	0.6903	0.6106	0.5328	0.7880
LAB [9]	1.4858	0.7636	0.8876	0.7048	**0.0533**	0.8030
Ours	**0.5262**	**0.8024**	**0.0000**	**0.7167**	0.1066	**0.8428**

Table 5: Failure rate and Area under curve(AUC) with NME threshold 0.2, normalized by inter-ocular distance, on Blurred-300VW dataset

Method	Category A		Category B		Category C	
	Failure Rate(%)	AUC	Failure Rate(%)	AUC	Failure Rate(%)	AUC
FAN [1]	14.5643	0.2659	4.1420	0.3853	0.9590	0.3975
RFLD [6]	25.7700	0.4965	15.5819	0.4347	3.0368	0.5885
SAN [2]	**3.4163**	0.5068	17.4556	0.3228	0.7987	0.5825
TCDCN [11]	13.8523	0.4627	17.5542	0.3120	1.0655	0.5805
LAB [9]	4.7361	0.5541	**2.1696**	0.4182	**0.5860**	0.6112
Ours	3.4670	**0.6293**	4.1420	**0.4585**	**0.5860**	**0.6909**

Table 6: Failure rate and Area under curve(AUC) with NME threshold 0.1, normalized by inter-ocular distance, on Blurred-300VW dataset

Method	Category A		Category B		Category C	
	Failure Rate(%)	AUC	Failure Rate(%)	AUC	Failure Rate(%)	AUC
FAN [1]	47.9183	0.2025	9.2702	0.2502	3.4630	0.2556
RFLD [6]	32.3015	0.4166	20.1183	0.3158	4.1556	0.4922
SAN [2]	10.6199	0.4244	33.5306	0.2114	0.9052	0.4786
TCDCN [11]	24.2687	0.3888	34.3195	0.2050	2.7171	0.4853
LAB [9]	8.0173	0.4635	9.9606	0.3037	**0.6926**	0.5147
Ours	**6.1136**	**0.5552**	**8.1854**	**0.3419**	0.7991	**0.6151**

Table 7: Failure rate and Area under curve(AUC) with NME threshold 0.08, normalized by inter-ocular distance, on Blurred-300VW dataset

Method	Category A		Category B		Category C	
	Failure Rate(%)	AUC	Failure Rate(%)	AUC	Failure Rate(%)	AUC
FAN [1]	3.7146	0.6394	0.8876	0.6999	**0.0533**	0.7135
RFLD [6]	13.0166	0.7022	7.0020	0.6968	1.4385	0.7978
SAN [2]	**0.6493**	0.7615	**0.0000**	0.6359	0.6390	0.8023
TCDCN [11]	1.2691	0.7039	0.3945	0.6305	0.4795	0.7995
LAB [9]	1.3775	0.7805	0.7890	0.7205	**0.0533**	0.8141
Ours	0.8667	**0.8170**	0.1972	**0.7325**	0.3729	**0.8508**

Table 8: Failure rate and Area under curve(AUC) with NME threshold 0.3, normalized by inter-pupil distance, on Blurred-300VW dataset

Method	Category A		Category B		Category C	
	Failure Rate(%)	AUC	Failure Rate(%)	AUC	Failure Rate(%)	AUC
FAN [1]	4.3337	0.4631	1.4793	0.5544	**0.0533**	0.5702
RFLD [6]	18.8980	0.6103	10.5523	0.5757	2.2376	0.7045
SAN [2]	1.2521	0.6477	3.5503	0.4732	0.6922	0.7039
TCDCN [11]	5.1850	0.5842	3.2544	0.4643	0.6926	0.7012
LAB [9]	2.2442	0.6785	1.0848	0.5835	0.2131	0.7226
Ours	**0.8667**	**0.7286**	**0.1972**	**0.6001**	0.3729	**0.7799**

Table 9: Failure rate and Area under curve(AUC) with NME threshold 0.2, normalized by inter-pupil distance, on Blurred-300VW dataset

Method	Category A		Category B		Category C	
	Failure Rate(%)	AUC	Failure Rate(%)	AUC	Failure Rate(%)	AUC
FAN [1]	63.0707	0.1863	18.2446	0.1725	10.3889	0.1729
RFLD [6]	36.4495	0.3790	27.5148	0.2533	5.5940	0.4322
SAN [2]	15.0873	0.3816	49.9014	0.1623	1.7572	0.4138
TCDCN [11]	31.3729	0.3600	48.8166	0.1445	6.0735	0.4338
LAB [9]	9.3329	0.4103	18.4418	0.2381	**0.7991**	0.4511
Ours	**7.7387**	**0.5151**	**12.4260**	**0.2694**	**0.7991**	**0.5643**

Table 10: Failure rate and Area under curve(AUC) with NME threshold 0.1, normalized by inter-pupil distance, on Blurred-300VW dataset

Method	Category A		Category B		Category C	
	Failure Rate(%)	AUC	Failure Rate(%)	AUC	Failure Rate(%)	AUC
FAN [1]	3.6682	0.7036	0.8876	0.7339	0.0533	0.7648
RFLD [6]	10.5711	0.7376	6.3116	0.7268	0.4262	0.8271
SAN [2]	0.5256	0.8049	**0.0000**	0.6772	0.0533	0.8338
TCDCN [11]	0.7274	0.7568	0.1972	0.6714	**0.0000**	0.8326
LAB [9]	1.0060	0.8179	0.5917	0.7507	0.0533	0.8475
Ours	**0.4179**	**0.8518**	**0.0000**	**0.7625**	**0.0000**	**0.8779**

Table 11: Failure rate and Area under curve(AUC) with NME threshold 0.1, normalized by diagonal length of the bounding box, on Blurred-300VW dataset

Method	Category A		Category B		Category C	
	Failure Rate(%)	AUC	Failure Rate(%)	AUC	Failure Rate(%)	AUC
FAN [1]	3.7456	0.6300	0.9862	0.6681	0.0533	0.7059
RFLD [6]	13.4345	0.6979	7.6923	0.6702	1.2254	0.7910
SAN [2]	0.7111	0.7577	**0.0000**	0.5965	0.5325	0.7962
TCDCN [11]	1.2227	0.7001	1.1834	0.5959	0.0533	0.7912
LAB [9]	1.4549	0.7765	0.8876	0.6910	0.0533	0.8094
Ours	**0.4798**	**0.8154**	**0.0000**	**0.7031**	**0.0000**	**0.8474**

Table 12: Failure rate and Area under curve(AUC) with NME threshold 0.08, normalized by diagonal length of the bounding box, on Blurred-300VW dataset

Method	Category A		Category B		Category C	
	Failure Rate(%)	AUC	Failure Rate(%)	AUC	Failure Rate(%)	AUC
FAN [1]	4.6897	0.4141	2.4655	0.4784	0.1598	0.5301
RFLD [6]	20.4148	0.5847	12.5247	0.5142	2.3442	0.6769
SAN [2]	1.5013	0.6193	7.5099	0.3977	0.6930	0.6758
TCDCN [11]	4.9837	0.5477	7.6923	0.3894	0.6926	0.6729
LAB [9]	2.5693	0.6529	**1.3807**	0.5090	**0.1066**	0.6955
Ours	**0.9596**	**0.7090**	1.5779	**0.5344**	0.1598	**0.7572**

Table 13: Failure rate and Area under curve(AUC) with NME threshold 0.05, normalized by diagonal length of the bounding box, on Blurred-300VW dataset

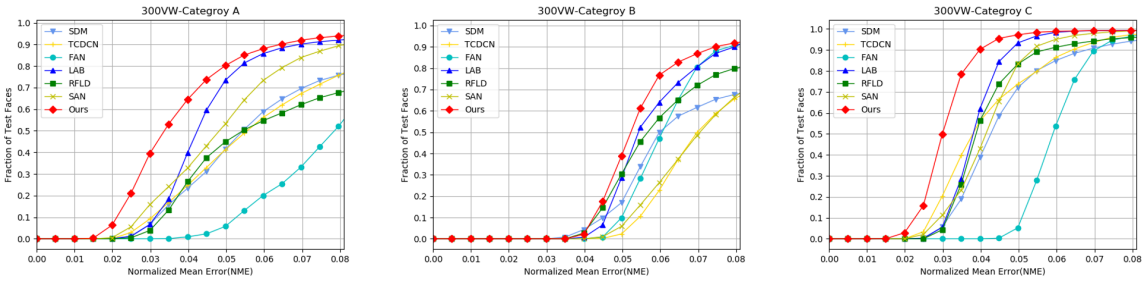

Figure 10: CED curves for testing results on Blurred-300VW, normalized by the distance between outer eye corner. Three sub-Figures from left to right corresponds to Category A, Category B and Category C.

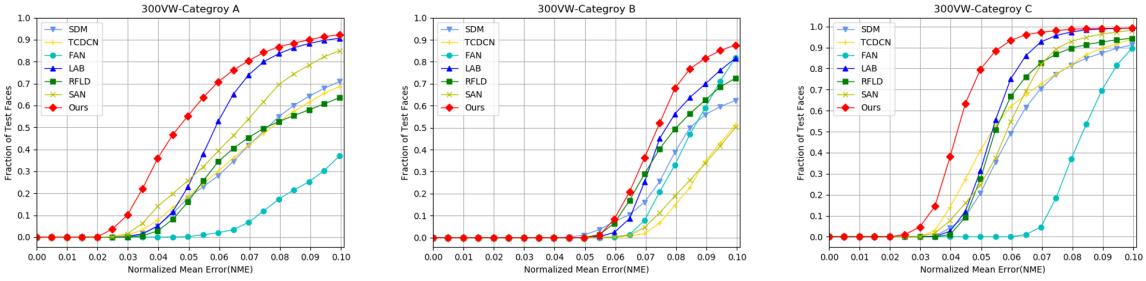

Figure 11: CED curves for testing results on Blurred-300VW, normalized by the distance between eye centers. Three sub-figures from left to right corresponds to Category A, Category B and Category C.

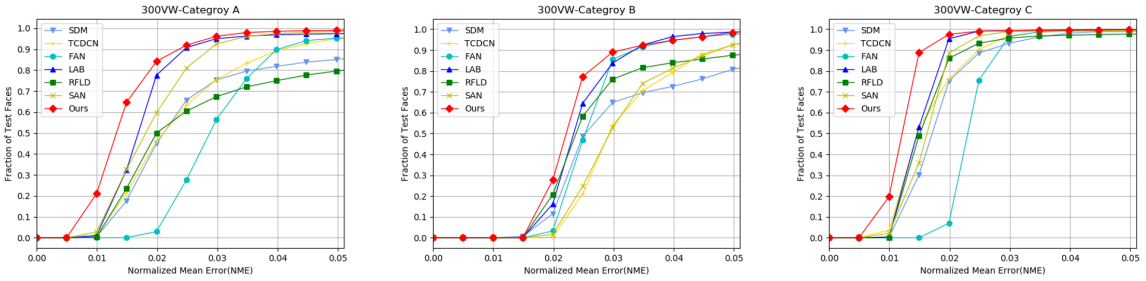

Figure 12: CED curves for testing results on Blurred-300VW, normalized by diagonal length of the bounding box. Three sub-figures from left to right corresponds to Category A, Category B and Category C.

Method	NME threshold: 0.2		NME threshold: 0.1		NME threshold: 0.08	
	Failure Rate(%)	AUC	Failure Rate(%)	AUC	Failure Rate(%)	AUC
FAN [1]	10.4162	0.5580	34.1836	0.2784	53.2606	0.1823
RFLD [6]	25.6669	0.5930	44.8399	0.3789	54.3852	0.2947
TCDCN [11]	11.2734	0.6045	34.8214	0.4154	45.7930	0.3474
SAN [2]	11.7909	**0.6276**	31.3739	0.4114	40.9356	0.3557
LAB [9]	8.2836	0.6093	32.1950	0.4254	41.1164	0.3426
Ours	**5.2462**	0.6272	**28.7715**	**0.4523**	**38.5429**	**0.3846**

Table 14: Failure rate and Area under curve(AUC), normalized by inter-ocular distance, on RWMB dataset

Method	NME threshold: 0.3		NME threshold: 0.2		NME threshold: 0.1	
	Failure Rate(%)	AUC	Failure Rate(%)	AUC	Failure Rate(%)	AUC
FAN [1]	9.3534	0.5833	17.9181	0.4357	65.0415	0.1399
RFLD [6]	24.1788	0.6109	32.6065	0.4967	60.3099	0.2562
TCDCN [11]	9.9362	0.6241	19.5159	0.5123	51.5600	0.3139
SAN [2]	10.4465	0.6451	18.6158	0.5370	46.0937	0.3159
LAB [9]	6.9876	0.6289	16.4644	0.5177	46.6365	0.3025
Ours	**4.1406**	**0.6467**	**12.8682**	**0.5372**	**43.7067**	**0.3496**

Table 15: Failure rate and Area under curve(AUC), normalized by inter-pupil distance, on RWMB dataset

Method	NME threshold: 0.1		NME threshold: 0.08		NME threshold: 0.05	
	Failure Rate(%)	AUC	Failure Rate(%)	AUC	Failure Rate(%)	AUC
FAN [1]	5.4447	0.6297	7.3442	0.5503	18.5490	0.3481
RFLD [6]	18.0964	0.6238	23.6645	0.5767	35.3700	0.4301
TCDCN [11]	3.8812	0.6479	8.0985	0.5910	22.5948	0.4537
SAN [2]	4.484	0.6687	8.6083	0.6148	20.5432	**0.4793**
LAB [9]	2.1189	0.6660	4.4092	0.5987	18.5764	0.4574
Ours	**0.4807**	**0.6902**	**1.5931**	**0.6207**	**13.5618**	0.4776

Table 16: Failure rate and Area under curve(AUC), normalized by diagonal length of the bounding box, on RWMB dataset

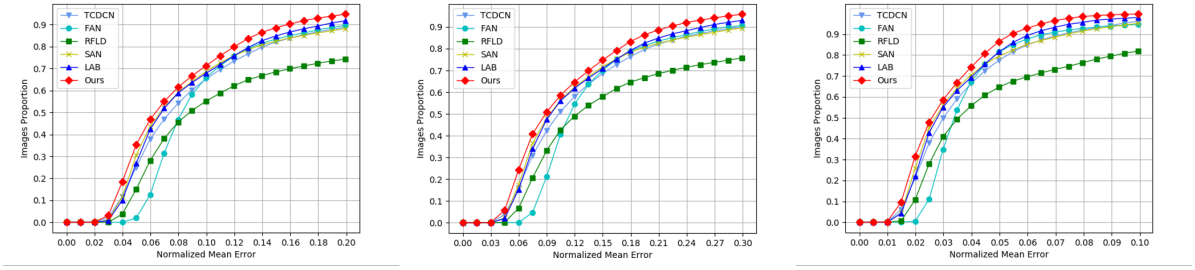

Figure 13: CED curves for testing results on RWMB data set. Three sub-figures from left to right corresponds to different normalization methods: by inter-ocular distance, by inter-pupil distance and by diagonal length of the bounding box

Figure 14: Deblurring effect and landmark accuracy comparison. Images in the first column are samples of Blurred-300VW. Landmarks of these blurry faces are presented in the second column. Red points are ground truth facial landmarks of the input and the green points are outputs of state-of-the-art facial landmark detection algorithms. By naively applying state-of-the-art deblurring and facial landmark detection algorithm, we got column three and column four. The proposed algorithm in the main paper produces the deblurred picture in column five and landmarks in column six.

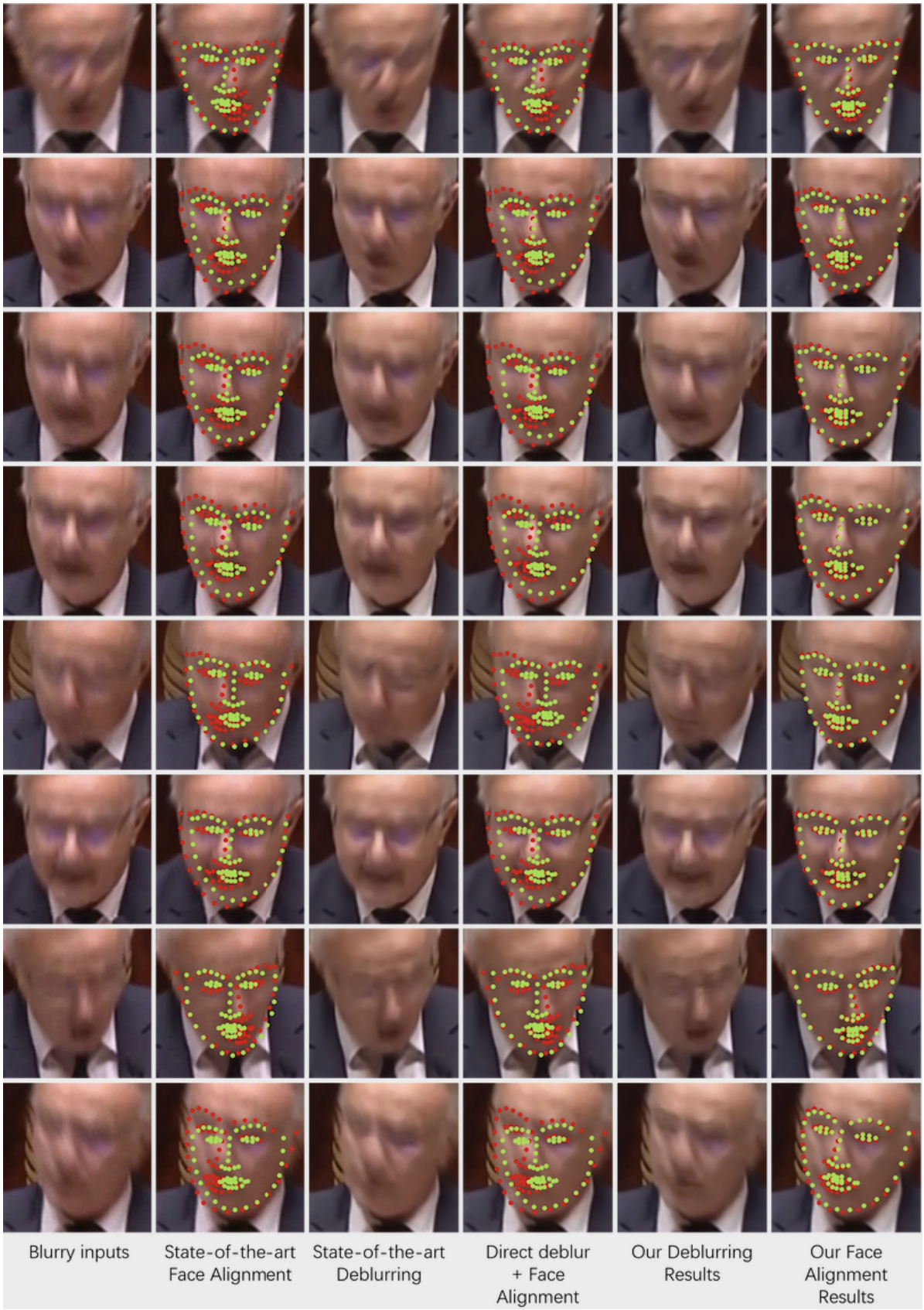

Figure 15: Deblurring effect and landmark accuracy comparison.

Figure 16: Annotation results of our RWMB dataset.

3. Benchmark: Real-World Motion Blur

Blurry faces in 300VW are not enough, less than in real-world circumstances, since the contents of videos in 300VW are mainly speeches and lectures. However, motion blur is common in realistic videos. Therefore, a dataset with severe motion blur is required.

As is mention in the main paper, we proposed a new benchmark named Real-World Motion Blur (RWMB). It contains 20 videos with obvious motion blur picked from YouTube, which include dancing, boxing, jumping, etc. The duration of each video is about one minute, thus each video contains approximately 1,800 frames. Hence, roughly 36,000 faces are annotated with 68 landmarks.

It is challenging to determine the specific location of each landmark. The annotation of previous frame is presented to the annotator as reference. Each frame is annotated by three expert annotators and checked by two quality inspectors.

We randomly picked one frame in each video and show the annotation results in Figure 16. Samples show that our annotation is accurate no matter the frame is clear or blurry.

4. Evaluate on realistic videos

In the YouTube, we presented the evaluation results of our algorithm on realistic videos, and compare it with other state-of-the-art facial landmark localization algorithms.

References

[1] 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 *ICCV*, 2017. 4, 6, 7, 8, 10

[2] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *CVPR*, 2018. 4, 6, 7, 8, 10

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016. 3

[5] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *ICCV*, 2017. 2

[6] Daniel Merget, Matthias Rock, and Gerhard Rigoll. Robust facial landmark detection via a fully-convolutional local-global context network. In *CVPR*, 2018. 4, 6, 7, 8, 10

[7] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *ICCV-W*, 2015. 4

[8] Patrick Wieschollek, Michael Hirsch, Bernhard Schölkopf, and Hendrik PA Lensch. Learning blind motion deblurring. In *ICCV*, 2017. 4

[9] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, 2018. 4, 6, 7, 8, 10

[10] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013. 4

[11] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014. 4, 6, 7, 8, 10