

Supplementary material.

Shape Preserving Facial Landmarks with Graph Attention Networks

Andrés Prados-Torreblanca^{1,2}

a.prados@upm.es

José M. Buenaposada¹

josemiguel.buenaposada@urjc.es

Luis Baumela²

lbaumela@fi.upm.es

¹ ETSII

Universidad Rey Juan Carlos
Móstoles, Spain

² Departamento de Inteligencia Artificial.

Universidad Politécnica de Madrid,
Boadilla del Monte, Spain

1 Implementation Details

In this section, we present a complete overview of SPIGA’s implementation. Including an extended study of the CNN multi-stage backbone configuration used to provide the initialization of the 2D landmark location and the visual feature representation (F) for our GAT regressor (see Fig. 1).

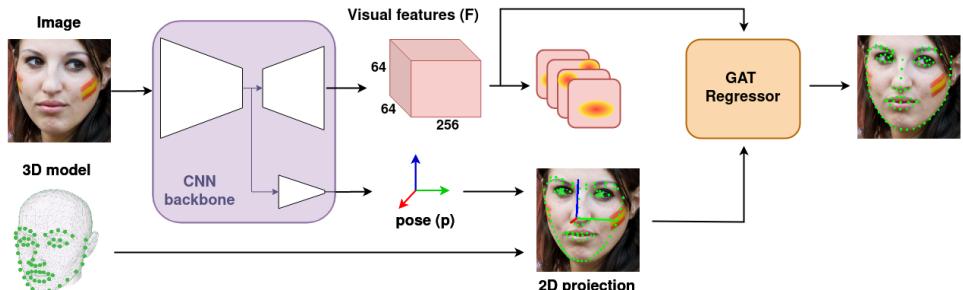


Figure 1: SPIGA workflow. Given as inputs an image and the facial 3D model, the CNN (MTN) infers the pose parameters, p , and the visual feature representation, F . Iteratively, the cascaded GAT regressor refines the initial 2D landmark projection provided by the 3D model, combining visual and structural information.

During training, we perform random data augmentation to the input images using the following transformations: rotation $\pm 45^\circ$, scaling $60 \pm 15\%$ of the bounding box size, translation 5% of the bounding box size, horizontal flip 50%, blur 50%, HSV color jittering and synthetic rectangular occlusions. Input face images are finally cropped and resized to

256×256 pixels. Similarly, 64×64 output heatmaps are generated following Awing [21] recommendations.

1.1 CNN Multitask Backbone

Our backbone (MTN) consists of a cascade of $M = 4$ Hourglass stages (HG) with an Attention Module, similar to the one used by [6]. First, a residual encoder reduces the size of the input image from 256×256 to 64×64 pixels before entering the HG cascade. Each HG reduces the spatial extent of the feature maps to a resolution of 8×8 at the bottleneck. Following [19], we add an encoder to each HG bottleneck to extract a 3D pose estimation head, as shown in Fig. 2.

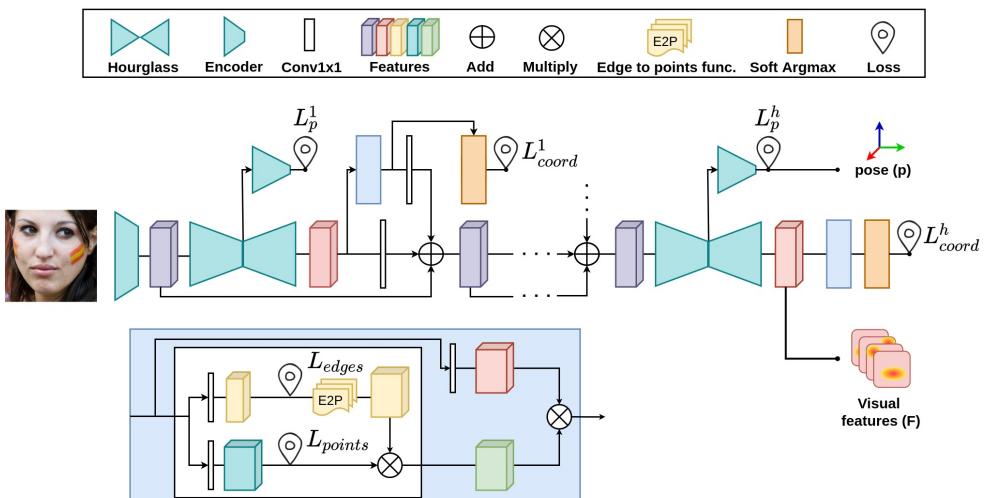


Figure 2: CNN Multitask backbone (MTL) architecture used during the fine-tuning with landmarks and pose estimation tasks.

We first pre-train the backbone in the landmark detection task (without the pose encoders) using the Adam optimizer during 450 epochs with an initial learning rate of 10^{-3} and a step decay of 0.1 at epoch 380. During training, the batch size is set to 24 and the Automatic Mixed Precision (AMP) from Pytorch is used. In Equation 1, we show the loss function computed for the landmark detection task. We aggregate the losses of each HG module, represented by index h , doubling the loss weight of a module compared to the previous one.

$$\mathcal{L}_{lnd} = \sum_{h=1}^M 2^{h-1} (\lambda_c \mathcal{L}_{coord}^h + \lambda_{att} (\mathcal{L}_{points}^h + \mathcal{L}_{edges}^h)), \quad (1)$$

Where λ_{coord} and λ_{att} are empirically set to 4 and 50, respectively. \mathcal{L}_{coord} is a smooth L1 function computed between the annotated and predicted landmarks coordinates. \mathcal{L}_{points} and \mathcal{L}_{edges} are Awing losses [21] applied to the point and edges heatmaps, respectively.

Once the model has been pre-trained with landmarks, it is fine-tuned with both tasks, pose and landmarks. Sharing the same hyperparameter configuration as in the previous pre-training stage during 150 epochs, with a step decay from 10^{-3} to 10^{-4} at epoch 100. In Equation 2, we show the final loss, where λ_p is empirically set to 1 and \mathcal{L}_{pose} is the L2 loss computed for the pose estimation. Once the model is trained, we freeze the backbone to train the GAT regressor.

$$\mathcal{L}_{total} = \mathcal{L}_{Ind} + \sum_{h=1}^M 2^{h-1} (\lambda_p \mathcal{L}_{pose}^h) \quad (2)$$

1.2 Cascaded Regressor Based on GATs

The full cascaded regressor is shown in Fig. 3 and the architecture of a single-step regressor is shown in Fig. 4. Similar to previous training configurations, the full shape regressor uses the Adam optimizer, setting an initial learning rate of 10^{-4} with a step decay of 0.1 at epoch 100.

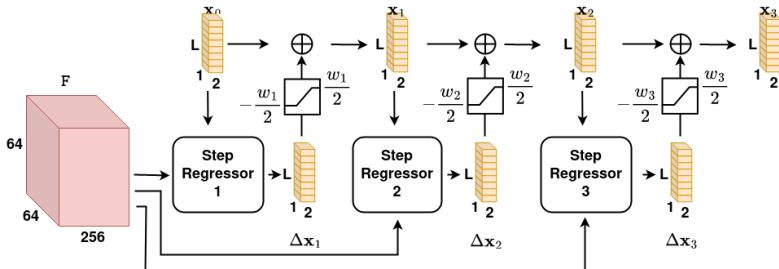


Figure 3: SPIGA cascaded regressor with the 3 steps used in the paper.

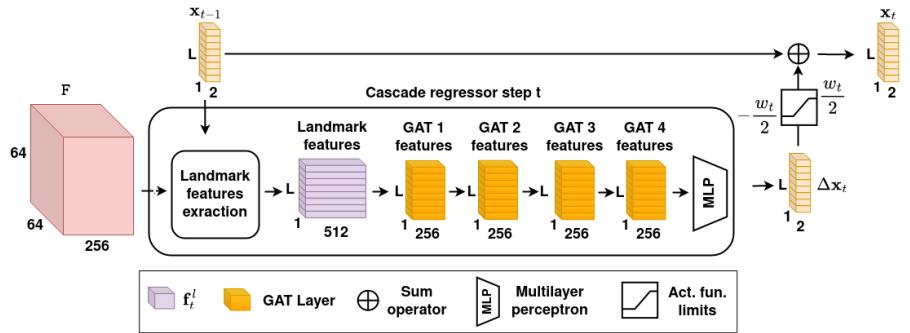


Figure 4: SPIGA step regressor with the 4 GATs layers used in the paper.

The detailed extraction of visual and geometric features can be visualized in Fig. 5. Including the encoding and combination applied to get the input features of the regressor.

Let F be the last feature map of the last stacked HG module in the MTN. We first look at a square window, \mathcal{W}_t , of size $w_t \times w_t$ and centered at each landmark location, \mathbf{x}_{t-1}^l in F . We use a fixed affine transform with the grid generator and sampler of the *Spatial Transformer Networks* [10] to have a differentiable crop operation of \mathcal{W}_t . The crop operation re-samples \mathcal{W}_t to a fixed size $7 \times 7 \times 256$ tensor, regardless of the dimension of the $w_t \times w_t$ window. Then, using a convolution with a 7×7 kernel, a $1 \times 1 \times 256$ feature map is extracted. Finally, with a 1×1 convolution, we compute the 512 channels of the visual features vector, \mathbf{v}_t^l , corresponding to l -th landmark at step t . For each landmark l , we combine the visual features extracted from the backbone network, \mathbf{v}_t^l , and the relative positional features, \mathbf{r}_t^l , computed from \mathbf{x}_{t-1} (i.e. the current shape) into the encoded features, $\mathbf{f}_t^l = \mathbf{v}_t^l + \mathbf{r}_t^l$.

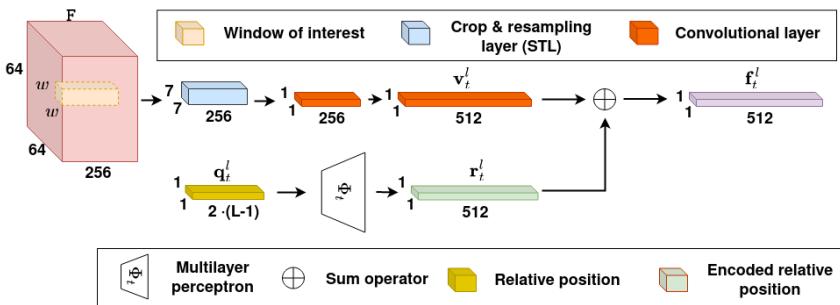


Figure 5: SPIGA extraction of visual and geometric features. Including the encoding and combination applied to get the input features of the regressor.

2 Extended Experimentation

In this section, we report an extended study of our proposal by adding new results on 300W (public and private) and WFLW datasets. In all our tables, results ranked **first**, **second** and **third** are shown respectively in blue, green and red colors.

300W public. In Table 1, we present the comparison of state-of-the-art (SOTA) results in the 300W public. In this dataset, our approach achieves results comparable to the top performers in the literature: ADNet [8] and SLD[13]. Since most images in this data set are fully visible semi-frontal faces, CNN-based methods already have a highly accurate performance (e.g. Wing). Our method is better than the other two methods using Graph Neural Networks (GraphNets), SDFL[12] and SLD[13], although results are comparable with SLD[13] ($NME_{int-ocul}$ of 2.99 vs 3.04). ADNet [8], using a stacked encoder-decoder model is the SOTA and our method obtains a comparable result ($NME_{int-ocul}$ of 2.93 vs 2.99).

300W private. Table 2 shows an extended SOTA comparison in terms of $NME_{int-ocul}$ on 300W private dataset.

WFLW. In Table 5 we present an extended SOTA comparison on WFLW.

Method	$NME_{int-ocul} (\%) (\downarrow)$			$NME_{int-pupil} (\%) (\downarrow)$		
	Common	Challeng.	Full	Common	Challeng.	Full
mnv2 [1]	3.93	7.52	4.70	-	-	-
SAN [2]	3.34	6.60	3.98	-	-	-
DAN [3]	3.19	5.24	3.59	4.42	7.57	5.03
TSR [4]	-	-	-	4.36	7.56	4.99
RAR [5]	-	-	-	4.12	8.35	4.94
LAB (4-stack) [6]	2.98	5.19	3.49	4.20	7.41	4.92
FTYM [7]	3.09	4.86	-	-	-	-
DeCaFA [8]	2.93	5.26	3.39	-	-	-
SHN [9]	-	4.90	-	4.12	7.00	4.68
HIIhc* [10]	2.95	5.04	3.36	-	-	-
HRNetV2-W18 [11]	2.87	5.15	3.32	-	-	-
HG×2+SAAT [12]	2.87	5.03	3.29	-	-	-
DCFE [13]	2.76	5.22	3.24	3.83	7.54	4.55
AVS [14]	-	-	-	3.98	7.21	4.54
PCD-CNN [15]	-	-	-	3.67	7.62	4.44
SDFL [16]	2.88	4.93	3.28	-	-	-
LUVLI [17]	2.76	5.16	3.23	-	-	-
SPLT [18]	2.75	4.90	3.17	-	-	-
3DDE [19]	2.69	4.92	3.13	3.73	7.10	4.39
GlomFace [20]	2.72	4.79	3.13	-	-	-
AWing [21]	2.72	4.52	3.07	3.77	6.52	4.31
SLD [22]	2.62	4.77	3.04	-	-	-
DTLD-s [23]	2.67	4.56	3.04	-	-	-
ADNet [24]	2.53	4.58	2.93	3.51	6.47	4.08
Wing [25]	-	-	-	3.27	7.18	4.04
SPIGA (Ours)	2.59	4.66	2.99	3.59	6.73	4.20

Table 1: Comparison against state-of-the-art on 300W public dataset.

Method	Indoor			Outdoor			Full		
	$NME_{inter-ocul}$ (\downarrow)	AUC_8 (\uparrow)	FR_8 (\downarrow)	$NME_{inter-ocul}$ (\downarrow)	AUC_8 (\uparrow)	FR_8 (\downarrow)	$NME_{inter-ocul}$ (\downarrow)	AUC_8 (\uparrow)	FR_8 (\downarrow)
DAN [26]	-	-	-	-	-	-	4.30	47.00	2.67
SHN [27]	4.10	-	-	4.00	-	-	4.05	-	-
DCFE [13]	3.96	52.28	2.33	3.81	52.56	1.33	3.88	52.42	1.83
3DDE [19]	3.74	53.93	2.00	3.71	53.95	2.66	3.73	53.94	2.33
SPIGA (Ours)	3.43	57.35	1.00	3.43	57.17	0.33	3.43	57.27	0.67

Table 2: Results on 300W private test set. Face alignment methods are exclusively trained on 300W public dataset.

3 Extended Ablation study

In this section we show more examples of the learned adjacency matrix per GAT module in the first cascade step (i.e. the attention of each landmark to others within the face graph). In Fig. 6 and Fig. 7 we show the estimated landmark locations (green dots) by SPIGA. On top of landmarks locations, we show as edges the attention estimated in the first cascade regressor step for two landmarks: one from the eye pupil (see Fig. 6) and one from the jaw (see Fig. 7). From left to right, we show the attention estimated in GAT 1 to 4.

When we have no occlusions (see the first row in Fig. 6) to estimate the pupil features, GAT 1 looks mainly at the other eye landmarks. Then, GATs progressively pay more attention to closer landmarks and also to the other pupil. To compute the pupil displacement, GAT 4 only attends the landmarks over the same eye. Interestingly, when we have the other eye occluded (see second and third rows in Fig. 6) GAT 1 does not pay only attention to the other eye landmarks, but it looks mainly to landmarks over the nose. Finally, when we have heavy occlusions (see the last row in Fig. 6), the attention is given first to not occluded parts (i.e. nose and the other eye in GAT 1) and to landmarks over the same eye in GAT 4.

Metric	Method	Testset	Pose	Expression	Illumination	Make-up	Occlusion	Blur
Bounding boxes from WFLW benchmark								
	mnv2 [1]	9.57	18.18	9.93	8.98	9.92	11.38	10.79
	LAB [2]	5.27	10.24	5.51	5.23	5.15	6.79	6.32
	SAN [3]	5.22	10.30	5.71	5.19	5.49	6.83	5.80
	Wing [4]	5.11	8.75	5.36	4.93	5.41	6.37	5.81
	3DDE [5]	4.68	8.62	5.21	4.65	4.60	5.77	5.41
	DeCaFa [6]	4.62	8.11	4.65	4.41	4.63	5.74	5.38
	AVS+SAN [7]	4.39	8.42	4.68	4.24	4.37	5.60	4.86
	AWing [8]	4.36	7.38	4.58	4.32	4.27	5.19	4.96
Bounding boxes from GT landmarks								
<i>NME_{ic} (%) (↓)</i>	GlomFace [9]	4.81	8.17	-	-	-	5.14	-
	HRNetV2-W18 [10]	4.60	7.94	4.85	4.55	4.29	5.44	5.42
	LUVLI [11]	4.37	7.56	4.77	4.30	4.33	5.29	4.94
	SDFL [12]	4.35	7.42	4.63	4.29	4.22	5.19	5.08
	AWing [13]	4.21	7.21	4.46	4.23	4.02	4.99	4.82
	SLD [14]	4.21	7.36	4.49	4.12	4.05	4.98	4.82
	HIHc [15]	4.18	7.20	4.19	4.45	3.97	5.00	4.81
	ADNet [16]	4.14	6.96	4.38	4.09	4.05	5.06	4.79
	DTLD-s [17]	4.14	-	-	-	-	-	-
	SPLT [18]	4.14	6.96	4.45	4.05	4.00	5.06	4.79
	SPIGA (Ours)	4.06	7.14	4.46	4.00	3.81	4.95	4.65
<i>FR₁₀ (%) (↓)</i>	HRNetV2-W18 [10]	4.64	23.01	3.50	4.72	2.43	8.29	6.34
	GlomFace [9]	3.77	17.48	-	-	-	6.73	-
	DTLD-s [17]	3.44	-	-	-	-	-	-
	LUVLI [11]	3.12	15.95	3.18	2.15	3.40	6.39	3.23
	SDFL [12]	2.72	12.88	1.59	2.58	2.43	5.71	3.62
	AWing [13]	2.04	9.20	1.27	2.01	0.97	4.21	2.72
	SLD [14]	3.04	15.95	2.86	2.72	1.46	5.29	4.01
	HIHc [15]	2.96	15.03	1.59	2.58	1.46	6.11	3.49
	ADNet [16]	2.72	12.72	2.15	2.44	1.94	5.79	3.54
	SPLT [18]	2.76	12.27	2.23	1.86	3.40	5.98	3.88
	SPIGA (ours)	2.08	11.66	2.23	1.58	1.46	4.48	2.20
<i>AUC₁₀ (%) (↑)</i>	HRNetV2-W18 [10]	52.37	25.06	51.02	53.26	54.45	45.85	45.15
	LUVLI [11]	57.70	31.00	54.90	58.40	58.80	50.50	52.50
	SDFL [12]	57.59	31.32	55.01	58.47	58.31	50.35	51.47
	AWing [13]	58.95	33.37	57.18	59.58	60.17	52.75	53.93
	SLD [14]	58.93	31.50	56.63	59.53	60.38	52.35	53.29
	HIHc ¹ [15]	59.70	34.20	59.00	60.60	60.40	52.70	54.90
	ADNet [16]	60.22	34.41	52.34	58.05	60.07	52.95	54.80
	SPLT [18]	59.50	34.80	57.40	60.10	60.50	51.50	53.50
	SPIGA (Ours)	60.56	35.31	57.97	61.31	62.24	53.31	55.31

Table 3: Extended evaluation of landmark detection on WFLW.

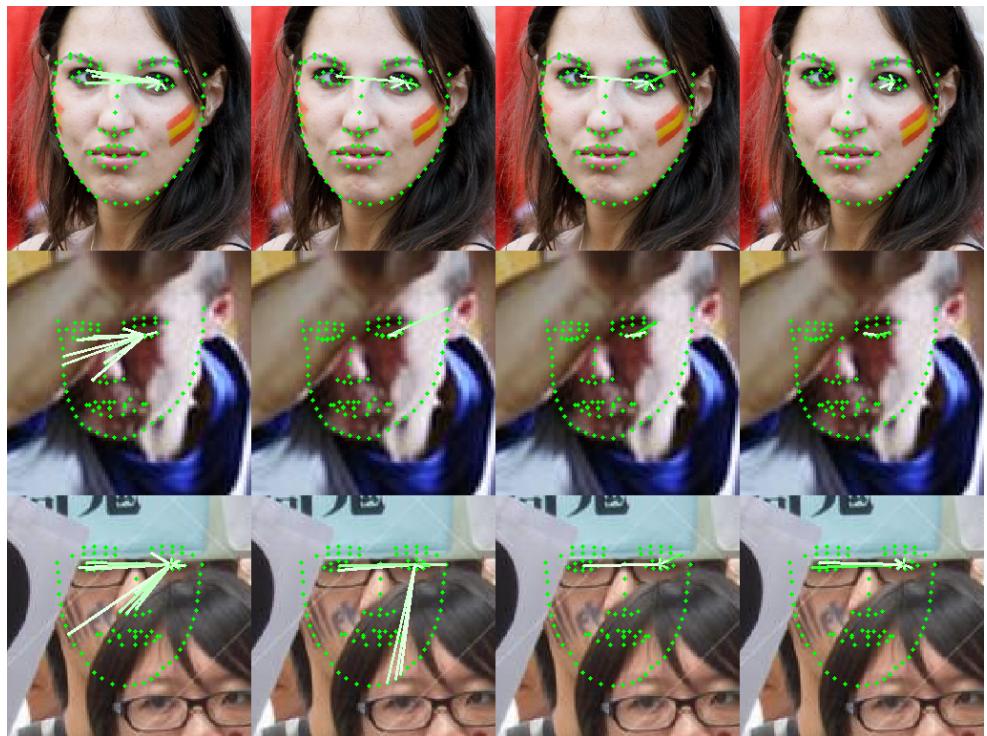


Figure 6: Attention from left eye pupil to other landmarks shown as edges. From left to right, attention at GAT layer 1, GAT layer 2, GAT layer 3 and GAT layer 4. The greener the higher is the attention. We only show edges with an attention over a threshold for clarity.

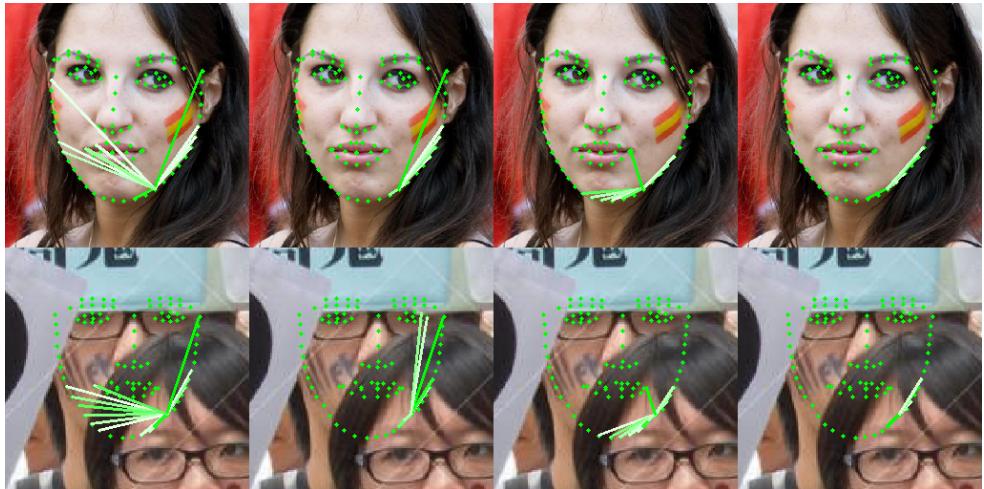


Figure 7: Attention from a landmark over the jaw to other landmarks shown as edges. From left to right, attention at GAT layer 1, GAT layer 2, GAT layer 3 and GAT layer 4. The greener the higher is the attention. We only show edges with an attention over a threshold for clarity.

Now we study the estimated attention of a jaw landmark (see Fig. 7). Without occlusions (first row in Fig. 7), the jaw landmark is paying attention to the mouth and other distant jaw landmarks in GAT 1. Progressively, the attention is concentrated on closer jaw landmarks. When we have heavy occlusions, the attention is given first to non-occluded landmarks in GAT 1. This allows the first graph convolution to compute features that use non-occluded landmarks. Then, the other GATs can use closer landmarks given that the starting features were free of occlusions.

We can conclude that the estimated attention allows us to extract occlusion-free features in the first GAT module. Then, the next GAT modules can use features from closer landmarks given the initial ones are correct.

4 Challenging examples

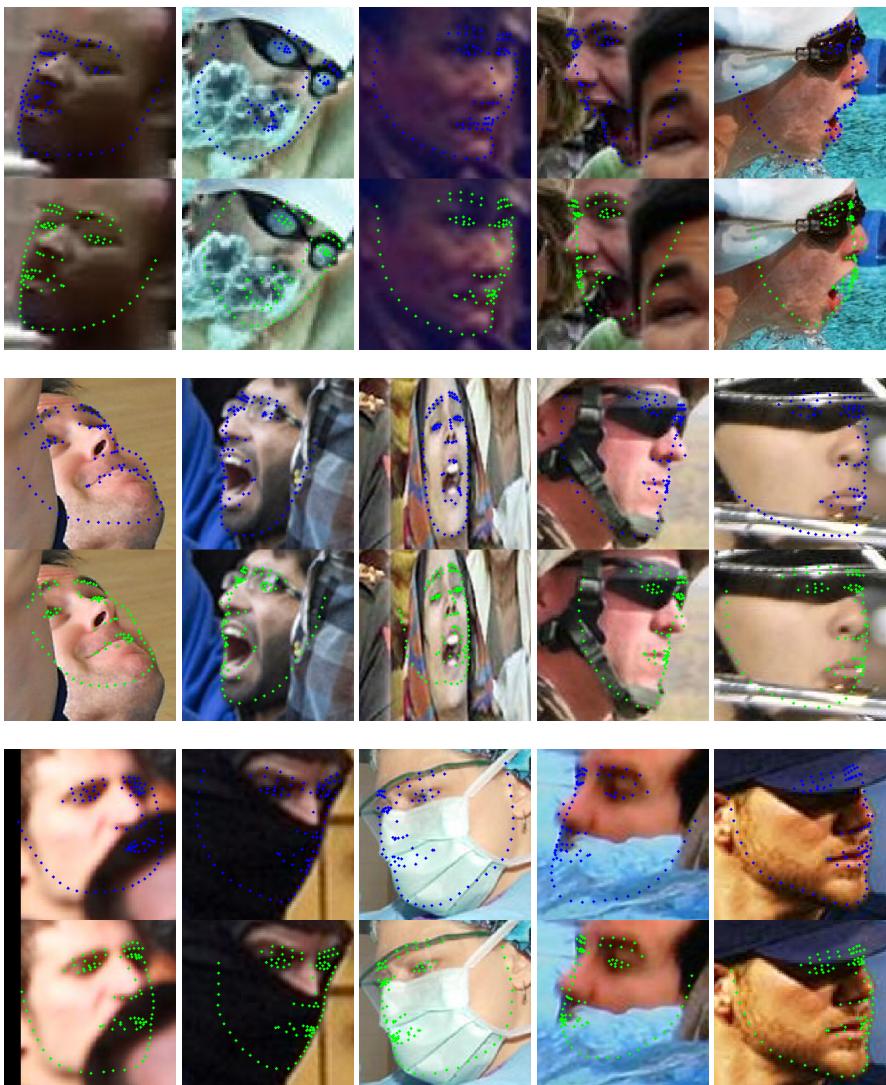


Figure 8: WFLW Challenging examples. In blue we show the ground truth and in green the landmark locations estimated by SPIGA.

References

- [1] Xiao Chu, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Structured feature learning for pose estimation. In *CVPR*, June 2016.
- [2] Arnaud Dapogny, Matthieu Cord, and Kevin Bailly. Decaf: Deep convolutional cascade for face alignment in the wild. In *ICCV*, pages 6892–6900. IEEE, 2019.
- [3] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *CVPR*, pages 379–388, 2018.
- [4] Ali Pourramezan Fard, Hojjat Abdollahi, and Mohammad H. Mahoor. Asmnet: A lightweight deep neural network for face alignment and pose estimation. In *CVPRW*, pages 1521–1530. CVF/IEEE, 2021.
- [5] 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 *CVPR*, pages 2235–2245, 2018.
- [6] Yangyu Huang, Hao Yang, Chong Li, Jongyoo Kim, and Fangyun Wei. Adnet: Leveraging error-bias towards normal direction in face alignment. In *ICCV*, pages 3080–3090, October 2021.
- [7] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *NeurIPS*, pages 2017–2025, 2015.
- [8] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *CVPRW*, pages 2034–2043, 2017.
- [9] 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 *CVPR*, pages 8233–8243, 2020.
- [10] Amit Kumar and Rama Chellappa. Disentangling 3D pose in a dendritic CNN for unconstrained 2D face alignment. In *CVPR*, pages 430–439, 2018.
- [11] Xing Lan, Qinghao Hu, and Jian Cheng. Revisting quantization error in face alignment. In *ICCVW*, pages 1521–1530, October 2021.
- [12] 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 CVPR*, pages 4176–4185, June 2022.
- [13] Weijian Li, Yuhang Lu, Kang Zheng, Haofu Liao, Chihung Lin, Jiebo Luo, Chi-Tung Cheng, Jing Xiao, Le Lu, Chang-Fu Kuo, and Shun Miao. Structured landmark detection via topology-adapting deep graph learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, pages 266–283. Springer International Publishing, 2020.

- [14] Chunze Lin, Beier Zhu, Quan Wang, Renjie Liao, Chen Qian, Jiwen Lu, and Jie Zhou. Structure-coherent deep feature learning for robust face alignment. *IEEE TIP*, 30: 5313–5326, 2021.
- [15] 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 *CVPR*, pages 3691–3700, 2017.
- [16] Shengju Qian, Keqiang Sun, Wayne Wu, Chen Qian, and Jiaya Jia. Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In *ICCV*, October 2019.
- [17] Roberto Valle, José M. Buenaposada, Antonio Valdés, and Luis Baumela. A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment. In *ECCV*, pages 609–624, 2018.
- [18] Roberto Valle, José M. Buenaposada, Antonio Valdés, and Luis Baumela. Face alignment using a 3D deeply-initialized ensemble of regression trees. *CVIU*, 189:102846, 2019.
- [19] Roberto Valle, José M. Buenaposada, and Luis Baumela. Multi-task head pose estimation in-the-wild. *IEEE TPAMI*, 43(8):2874–2881, 2021.
- [20] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, 43(10):3349–3364, 2021.
- [21] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *ICCV*, October 2019.
- [22] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Sebastian Dziadzio, Matthew Johnson, Virginia Estellers, Tom Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone. In *ICCV*, October 2021.
- [23] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, pages 2129–2138, 2018.
- [24] 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 CVPR*, pages 4052–4061, June 2022.
- [25] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf A. Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *ECCV*, pages 57–72, 2016.
- [26] Jing Yang, Qingshan Liu, and Kaihua Zhang. Stacked hourglass network for robust facial landmark localisation. In *CVPRW*, pages 2025–2033, 2017.
- [27] Congcong Zhu, Xiaoqiang Li, Jide Li, and Songmin Dai. Improving robustness of facial landmark detection by defending against adversarial attacks. In *ICCV*, pages 11751–11760, October 2021.

- [28] Congcong Zhu, Xintong Wan, Shaorong Xie, Xiaoqiang Li, and Yinzhen Gu. Occlusion-robust face alignment using a viewpoint-invariant hierarchical network architecture. In *Proceedings of the IEEE/CVF CVPR*, pages 11112–11121, June 2022.