Shape Preserving Facial Landmarks with Graph Attention Networks

Andrés Prados-Torreblanca[1,2]
a.prados@upm.es
José M. Buenaposada[1]
josemiguel.buenaposada@urjc.es
Luis Baumela[2]
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

Abstract

Top-performing landmark estimation algorithms are based on exploiting the excellent ability of large convolutional neural networks (CNNs) to represent local appearance. However, it is well known that they can only learn weak spatial relationships. To address this problem, we propose a model based on the combination of a CNN with a cascade of Graph Attention Network regressors. To this end, we introduce an encoding that jointly represents the appearance and location of facial landmarks and an attention mechanism to weigh the information according to its reliability. This is combined with a multi-task approach to initialize the location of graph nodes and a coarse-to-fine landmark description scheme. Our experiments confirm that the proposed model learns a global representation of the structure of the face, achieving top performance in popular benchmarks on head pose and landmark estimation. The improvement provided by our model is most significant in situations involving large changes in the local appearance of landmarks. The code is publicly available at https://github.com/andresprados/SPIGA

1 Introduction

Landmarks (or keypoints) are a widely used representation to address high-level vision tasks such as image retrieval [18], facial expression recognition [23], face reenactment [35], etc. The performance of computer vision algorithms on the final task depends, to a great extent, on the accuracy and robustness of this intermediate representation. Thus, although many algorithms with excellent performance have recently emerged, research is still very intense in this area.

Top facial landmark estimation methods may be broadly grouped into coordinate and heatmap regression approaches. *Coordinate regression approaches* directly estimate the landmark position by projecting the representation estimated by a CNN encoder onto a set of 2D coordinates [6, 7, 12, 17, 24]. They are the most efficient since they only require an encoder architecture to compute the facial representation. The *heatmap regression approach* is based on appending multiple encoder-decoder modules to estimate a 2D data structure modeling the landmark position likelihood, the heatmap [8, 9, 10, 13, 30, 31]. The landmark

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

coordinates are typically estimated at the maximum of each heatmap. This architecture provides an increase in accuracy at the expense of a considerable boost in computational and memory requirements. A fundamental limitation of both approaches is their degradation when there is ambiguity or noise contaminating the local landmark appearance. This typically happens at the presence of occlusions, heavy make-up, blur and extreme illuminations or poses. This is because of the known fact that CNNs cannot learn simple spatial relationships [21] and, in the case of facial landmarks, are unable to learn a global representation of the face structure. However, a human face is a highly structured object with a prominent landmark configuration. Therefore, an effective way of representing the local appearance of each landmark and its geometric relationship to the other landmarks is needed.

This problem has been partially addressed in the literature with a local attention module combining landmarks with facial boundaries [9, 10, 31]. This is a solution that learns short-distance geometrical relationships. An alternative solution combines the advantages of a CNN description with traditional Ensemble of Regression Trees (ERT) [25, 26]. Although this solution is able to learn long-distance geometrical dependencies, it is not fully satisfactory because of the limited learning capabilities of ERTs and the impossibility of end-to-end training. Other approaches use a Graph Convolutional Network (GCN) to learn the facial geometrical structure [16, 17]. This is achieved by combining the landmark local description, extracted from the CNN representation, with geometrical information represented by the relative landmark locations. However, poor initialization and the lack of an advanced attention mechanism reduce the performance of these models. More recent approaches use transformers [15, 32] in a cascade shape regressor, obtaining very good results due to the built-in attention mechanisms.

In this paper, we present the SPIGA (*Shape Preserving wIth GAts*) model for the estimation of human face landmarks. We follow the traditional regressor cascade approach [2] and present an algorithm that combines a multi-stage heatmap backbone with a cascade of Graph Attention Network (GAT) regressors [28]. The backbone provides a top-performing facial appearance representation. The cascaded GAT regressor is endowed with a positional encoding and attention mechanism that learn the geometrical relationship among landmarks. Another element of our proposal that improves the convergence of the GAT cascade is a coarse-to-fine feature extraction procedure and a good initialization. To do this, we train our backbone with a multi-task approach that also estimates the head pose, using its projection to establish the initial landmark locations. We evaluate the performance of our proposal in 300W, COFW-68, MERL-RAV and WFLW datasets. It achieves top performance on both head pose and face landmarks estimation. The improvement is most significant in situations involving large appearance changes, such as occlusions, heavy make-up, blur and extreme illuminations. We make the following contributions: 1) A GAT cascade with an attention mechanism to weigh the information provided by each landmark according to its reliability; 2) A positional encoding to jointly represent relative landmark locations and local appearance; 3) A multi-task approach to initialize the location of graph nodes; 4) A coarse-to-fine landmark description scheme.

2 Shape Regressor Model

We propose a coarse-to-fine cascade of landmark regressors [2, 4] that iteratively refines the landmarks coordinates while preserving the face shape. Our approach involves three critical components: 1) the initialization, 2) the features used for regression, and 3) the regressors

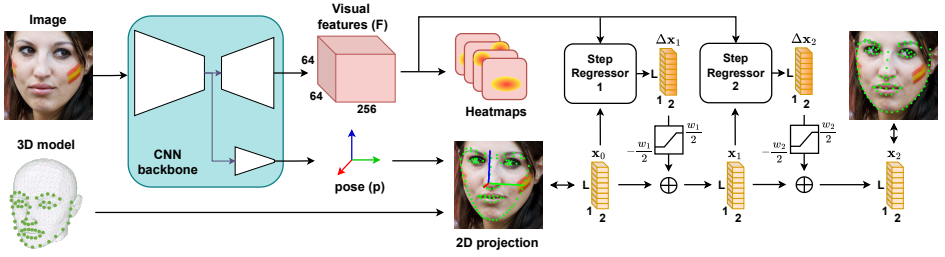

Figure 1: Regressor architecture with a two-step cascade.

that estimate the face shape deformation at each step of the cascade.

In our proposal, we use a multi-task CNN backbone to provide both, the initialization and the local appearance representation. We set the initial shape of the face, $\mathbf{x}_0 \in \mathbb{R}^{L\times 2}$, by projecting L landmarks from a generic 3D rigid face mesh oriented using the head pose backbone prediction. At each cascade step t, a GAT-based [28] regressor computes a displacement vector, $\Delta\mathbf{x}_t$, to update the landmarks location, $\mathbf{x}_t = \mathbf{x}_{t-1} + \Delta\mathbf{x}_t$. After K steps, the final face shape is $\mathbf{x}_K = \mathbf{x}_0 + \sum_{t=1}^{K} \Delta\mathbf{x}_t$. We denote the 2D location of l-th landmark at step t as $\mathbf{x}_t^l \in \mathbb{R}^2$. In Fig. 1 we show the regressor with a two-step cascade configuration.

2.1 Initialization by Head Pose Estimation

Our multi-task backbone, termed *Multi Task Network* (*MTN*), is a cascade of M encoder-decoder Hourglass (HG) modules. Each HG module in MTN is composed of a shared encoder with two task branches: 1) a 3D head pose estimation branch and 2) a landmark estimation decoder to the end of which we attach the next HG module. Defining and balancing the depth of the three components is a critical factor to boost the head pose estimation accuracy. We supervise the h-th module pose head by comparing its estimation, $\mathbf{p} \in \mathbb{R}^6$, with the ground truth, $\tilde{\mathbf{p}}$, using the L2 loss, $\mathcal{L}_{\mathbf{p}}^h(\mathbf{p},\tilde{\mathbf{p}}) = ||\tilde{\mathbf{p}} - \mathbf{p}||^2$. Our annotations for pose, $\tilde{\mathbf{p}}$, are obtained from the ground truth landmarks using a rigid head model (see Fig. 1). In the landmarks task we optimize a coordinate smooth L1 loss ($\mathcal{L}_{coord}$) enhanced by a local attention mechanism ($\mathcal{L}_{att}$) on the heatmaps, like [10, 30]. The final landmark loss is defined as $\mathcal{L}_{lnd} = \sum_{h=1}^{M} 2^{h-1}(\lambda_c \mathcal{L}_{coord}^h + \lambda_{att}\mathcal{L}_{att}^h)$, where λ's are scalars empirically optimized. For further details, please see the supplementary material.

To obtain a top-performing head pose estimation model (see Table 1) we pre-train the network only with the landmark task, $\mathcal{L}_{lnd}$, and fine-tune with both tasks, landmarks and pose, like [27]. For multi-task fine-tuning we use the loss $\mathcal{L}_{mt} = \mathcal{L}_{lnd} + \lambda_{\mathbf{p}} \sum_{h=1}^{M} 2^{h-1}\mathcal{L}_{\mathbf{p}}^h$, where $\lambda_{\mathbf{p}}$ is a hyperparameter. Although we use intermediate supervision at every HG module, the prediction of $\mathbf{p}$ to estimate $\mathbf{x}_0$, as well as the visual features, are extracted from the last module. Let $\mathbf{X} \in \mathbb{R}^{L\times 3}$ be the 3D coordinates on the 3D head model that correspond to the L 2D landmarks. If the pose estimated by the backbone is given by $\mathbf{p}$, then the *initial shape*, $\mathbf{x}_0$, is computed by projecting the 3D model, $\mathbf{x}_0 = \pi(\mathbf{X};\mathbf{p})$, where $\pi(\cdot)$ is the 3D$\rightarrow$2D projection function.

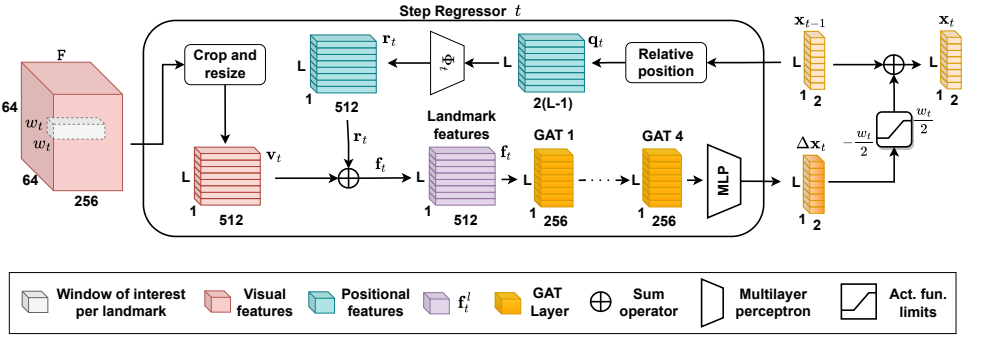

Figure 2: Appearance and shape feature extraction for the t-th step regressor.

2.2 Geometric and Visual Feature Extraction

For each step in the cascaded regressor, the input features are a combination of local appearance at each landmark (i.e. visual features) and global representation of the facial structure (i.e. geometric features). How visual and positional information is extracted and combined has a direct impact on the performance of the regressor (see Table 5).

Let F be the output feature map of the last stacked HG module in the MTN. We extract local appearance information from a square window, $\mathcal{W}_t$, of size $w_t \times w_t$, centered at each landmark location, $\mathbf{x}_{t-1}^l$, in F. We use a fixed affine transform with a grid generator and sampler [11] to crop and re-sample $\mathcal{W}_t$ at a fixed size, regardless of w_t. Then, using convolutional layers, we extract the visual features, $\mathbf{v}_t^l$, corresponding to the l-th landmark at step t. We iteratively reduce w_t at each step t, in a coarse-to-fine approach.

Positional information is crucial to maintain the shape of the face when local appearance alone is not sufficient (e.g. in presence of occlusions, blur, make-up, etc.). Relative distances between landmarks provide enhanced geometrical features compared to their absolute locations since they explicitly represent the facial shape. This relative positional information can be defined from displacement vectors between landmarks [16]. Let $\mathbf{q}_t^l = \{\mathbf{x}_{t-1}^l - \mathbf{x}_{t-1}^i\}_{i \neq l} \in \mathbb{R}^{2\times(L-1)}$ be the displacement vector corresponding to l-th landmark in the t-th step. In contrast to [16], we learn a high dimensional embedding from $\mathbf{q}_t^l$ using a Multi layer Perceptron (MLP), $\mathbf{r}_t^l = \Phi_t(\mathbf{q}_t^l)$, that facilitates the aggregation of the visual local appearance and the facial shape information. In the experiments, we show that this way of encoding relative positional information in $\mathbf{r}$ improves the shape-preserving ability of the network (see section 3.4).

Let $\mathbf{f}_t^l$ be the feature vector used to compute $\Delta\mathbf{x}_t^l$. At each step t of the cascade (see Fig. 2), and for each landmark l, we add the visual features extracted from the backbone network, $\mathbf{v}_t^l$, with the relative positional features, $\mathbf{r}_t^l$, computed from the current shape, $\mathbf{x}_{t-1}$, to produce the encoded features, $\mathbf{f}_t^l = \mathbf{v}_t^l + \mathbf{r}_t^l$.

2.3 Cascade Shape Regressor Using GATs

The step regressor architecture (Fig. 2) is composed of stacked GAT layers inspired by the ones in the Attentional Graph Neural Net [22]. We consider the facial shape as a single densely connected graph where nodes are the landmark locations, $\mathbf{x}_t$. To weigh the shared

information across nodes, we compute a dynamic adjacency matrix per GAT layer s, $\mathcal{A}_t^s$. We learn these matrices as an attention from a given landmark to every other in the graph.

The input to the first GAT layer at step t are the encoded features, $\{\mathbf{f}_t^i\}_{i=1}^L$. Let $\mathbf{f}_t^{i,s-1}$ be the features of the i-th landmark produced by the (s-1)-th GAT layer, that are also the input to s-th layer ($\mathbf{f}_t^{i,0} \equiv \mathbf{f}_t^i$). From now on, we drop the step-index t to simplify the notation. The updated feature vector after the s-th layer is defined as $\mathbf{f}^{i,s} = \mathbf{f}^{i,s-1} + MLP([\mathbf{f}^{i,s-1}||\mathbf{m}^{i,s}])$ where $[\cdot||\cdot]$ is the concatenation operator, $\mathbf{m}^{i,s}$ is the information aggregated, or message, of the nodes neighboring i. Focusing on the message generation procedure, a query vector $\mathbf{h}_q^{i,s}$, is assigned to landmark i and key $\mathbf{h}_k^{j,s}$, and value vectors $\mathbf{h}_v^{j,s}$, to every other landmark j. The attention weight of landmark i to landmark j is the `SoftMax` over the key-query similarities $\alpha_{ij} = \text{SoftMax}_j(\mathbf{h}_q^{i,s} \cdot \mathbf{h}_k^{j,s})$, being α_{ij} the elements of the adjacency matrix $\mathcal{A}_t^s$ and the transmitted message $\mathbf{m}^{i,s}$ the weighted average of the value vectors: $\mathbf{m}^{i,s} = \sum_{i \neq j} \alpha_{ij} \mathbf{h}_v^{j,s}$, where $\mathbf{h}_q^{i,s} = \mathtt{W}_1^s \mathbf{f}^{i,s} + \mathbf{b}_1^s$, $\mathbf{h}_k^{j,s} = \mathtt{W}_2^s \mathbf{f}^{j,s} + \mathbf{b}_2^s$ and $\mathbf{h}_v^{j,s} = \mathtt{W}_3^s \mathbf{f}^{j,s} + \mathbf{b}_3^s$. Matrices $\mathtt{W}_i$ and bias vectors $\mathbf{b}_i$ are learned.

Finally, the last GAT layer output $\mathbf{f}_t^{i,4}$ is processed by a decoder, an MLP, to obtain the corresponding displacement, $\Delta\mathbf{x}_t^i$. We constraint the values in $\Delta\mathbf{x}_t^i$, applying an `ArcTan` activation and scaling the result, to be in the interval $[-w_t/2, w_t/2]$. In practice, this constraint makes the single-step regressor search problem simpler, boosting training convergence. Given a trained MTN backbone, we train the cascade with the $\mathcal{L}_{CR} = \sum_{t=1}^K L1_{smooth}[\tilde{\mathbf{x}} - (\mathbf{x}_{t-1} + \Delta\mathbf{x}_t)]$ loss, where $\tilde{\mathbf{x}}$ are the ground truth landmark coordinates.

3 Experiments

To train and evaluate our method, we conduct different experiments in four complementary datasets which have been acquired in-the-wild and bear different levels of difficulty:

300W [20] provides 68 manually annotated landmarks. We employ the 300W private extension, which uses 3837 images as training set and adds 600 test images divided into indoor and outdoor subgroups.

COFW-68 is a re-annotated version of COFW [1] with 68 landmarks. It is conceived for testing landmark detectors with occlusions in a cross-dataset approach. The testing set in COFW-68 is made of 507 images. The annotations include the landmark positions and the visibility labels for the same 68 points as in 300W.

WFLW [31] is composed of challenging in-the-wild images and provides 98 manually annotated landmarks. The dataset has 7500 training and 2500 testing faces. It is divided into 6 subgroups: pose, expression, illumination, make-up, occlusion and blur.

MERL-RAV [13] is a re-annotated version of 19,000 AFLW images with 68 landmarks, like 300W. It provides 15,449 training and 3,865 test faces divided into 3 orientation subsets: frontal, half-profile and profile. This recent dataset includes externally occluded visibility and self-occluded labels.

3.1 Evaluation Metrics

In order to quantify the head pose estimation error, we use the Mean Absolute Error (MAE) metric, $MAE = \frac{1}{N}\sum_{i=1}^{N} |\tilde{p}_i - p_i|$, where N is the number of testing images, $\tilde{p}_i$ is the ground truth and p_i represents a single predicted pose parameter.

	300W Angular error (°)(↓)				WFLW Angular error (°)(↓)				MERL-RAV Angular error (°)(↓)			
Method	yaw	pitch	roll	mean	yaw	pitch	roll	mean	yaw	pitch	roll	mean
Yang [32]	4.2	5.1	2.4	3.9	-	-	-	-	-	-	-	-
JFA [33]	2.5	3.0	2.6	2.7	-	-	-	-	-	-	-	-
ASMNet [6]	1.62	1.80	1.24	1.55	2.97	2.93	2.21	2.70	-	-	-	-
MNN [27]	-	-	-	1.56	-	-	-	2.08	-	-	-	-
SPIGA (Ours)	**1.41**	**1.70**	**0.77**	**1.29**	**1.78**	**1.86**	**0.93**	**1.52**	**3.23**	**2.24**	**1.71**	**2.39**

Table 1: Head pose MAE, in degrees, for 300W public, WFLW and MERL-RAV datasets.

Focusing on the landmark estimation task, Normalized Mean Error (NME) is the standard metric, $NME = \frac{100}{N}\sum_{i=1}^{N}\sum_{l=1}^{L}\frac{||\tilde{\mathbf{x}}_l^i - \mathbf{x}_l^i||_2}{d_i}$. Where $\tilde{\mathbf{x}}_l^i$ and $\mathbf{x}_l^i$ denote, respectively, the ground-truth and predicted coordinates of the i-th landmark and d_i is a normalization value which varies depending on the dataset: inter-ocular (int-ocul), distance between outer eye corners; inter-pupils, distance between pupil/eye centers; and box, computed as the geometric mean of the landmarks ground truth bounding box ($d = \sqrt{w_{bbox} * h_{bbox}}$).

We also use Failure Rate (FR) and Area Under the Curve (AUC). FR evaluates the robustness of algorithms in terms of NME, indicating the percentage of images with an NME above a given threshold. AUC is calculated by computing the area under the Cumulative Error Distribution (CED) curve from 0 to the FR threshold. We introduce the Normalized mean Percentile Error 90 (NPE_{90}) which represents the NME for the image at the 90% of the dataset, sorted by NME. This metric is particularly convenient for small data subsets where the FR is not representative.

In all our tables results ranked **first**, second and third are shown respectively in blue, green and red colors.

3.2 3D Pose Estimation Results

First, we evaluate the MTN performance in 3D pose estimation. In Table 1, we compare our pose estimation in 300W and WFLW with previous works in the literature. Our model shows a significant improvement. We reduce the mean MAE of the previous top performer, MNN [27], by 17% and 27% respectively in 300W and WFLW. The main reason behind this improvement is a better network architecture, stacked HGs vs. a single encoder-decoder in [27] and the use of an attention mechanism. Having such a precise head pose estimation is a critical factor in our proposal, since the cascade shape regressor initialization relies on this prediction.

3.3 Landmark Detection Results

WFLW is the most popular benchmark to evaluate the performance of facial landmark detection. Recent methods that adopt this dataset use the bounding boxes provided by HRnet [29], that were obtained from the ground truth landmark annotations. By doing so, they achieve better performance (see Table 2, AWing results improve from 4.36 to 4.21 NME). In Table 2, we clearly distinguish the bounding boxes used in the evaluation. Another important aspect to perform a fair comparison is the use of additional training data. In our discussion we do not consider methods that train with images or annotations other than those provided by WFLW.

Metric	Method	Testset	Pose	Expression	Illumination	Make-up	Occlusion	Blur
$NME_{int-ocul}$ (%)(↓)	Bounding boxes from WFLW benchmark							
	3DDE [26]	4.68	8.62	5.21	4.65	4.60	5.77	5.41
	DeCaFA [8]	4.62	8.11	4.65	4.41	4.63	5.74	5.38
	AVS+SAN [15]	4.39	8.42	4.68	4.24	4.37	5.60	4.86
	AWing [30]	4.36	7.38	4.58	4.32	4.27	5.19	4.96
	Bounding boxes from GT landmarks							
	GlomFace [20]	4.81	8.17	-	-	-	5.14	-
	LUVLI [25]	4.37	7.56	4.77	4.30	4.33	5.29	4.94
	SDFL [27]	4.35	7.42	4.63	4.29	4.22	5.19	5.08
	AWing [30]	4.21	7.21	4.46	4.23	4.02	4.99	4.82
	SLD [16]	4.21	7.36	4.49	4.12	4.05	4.98	4.82
	HIHc[1] [24]	4.18	7.20	**4.19**	4.45	3.97	5.00	4.81
	ADNet [10]	4.14	**6.96**	4.38	4.09	4.05	5.06	4.79
	DTLD-s [15]	4.14	-	-	-	-	-	-
	SPLT [32]	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**
FR_{10} (%)(↓)	GlomFace [20]	3.77	17.48	-	-	-	6.73	-
	DTLD-s [15]	3.44	-	-	-	-	-	-
	LUVLI [25]	3.12	15.95	3.18	2.15	3.40	6.39	3.23
	SDFL [27]	2.72	12.88	1.59	2.58	2.43	5.71	3.62
	AWing [30]	**2.04**	**9.20**	**1.27**	2.01	**0.97**	**4.21**	2.72
	SLD [16]	3.04	15.95	2.86	2.72	1.46	5.29	4.01
	HIHc[1] [24]	2.96	15.03	1.59	2.58	1.46	6.11	3.49
	ADNet [10]	2.72	12.72	2.15	2.44	1.94	5.79	3.54
	SPLT [32]	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**
AUC_{10} (%)(↑)	AWing [30]	58.95	33.37	57.18	59.58	60.17	52.75	53.93
	SLD [16]	58.93	31.50	56.63	59.53	60.38	52.35	53.29
	HIHc[1] [24]	59.70	34.20	**59.00**	60.60	60.40	52.70	54.90
	ADNet [10]	60.22	34.41	52.34	58.05	60.07	52.95	54.80
	SPLT [32]	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 2: Evaluation of landmark detection on WFLW.

In Table 2, we show that our model outperforms current state-of-the-art (SOTA) in most of the WFLW subsets, as well as in the full set metrics. When it is compared with other GraphNets-based methods, our approach is 4% and 32% better in terms of NME and FR than SLD [16], and 7% and 23% better than SDFL [27]. These results show that our relative positional encoding and the per layer graph attention mechanism have a strong impact on the performance of GraphNets. Further, our proposal is also more accurate than recent approaches based on transformers, when these models are trained only with WFLW data, DTLD-s [15] and SPLT [32], both with 4.14 NME in the full set. If we analyze the performance on some of the subsets, our method is 35%, 25%, 23% and 39% better than the previous SOTA, ADNet [10], in the illumination, make-up, occlusion and blur subsets. This proves the importance of learning a global representation of the facial structure, that CNNs alone do not provide. Additionally, the low FR across the different subsets and better AUC values reaffirm that our model achieves a balanced trade-off between robustness and precision, taking advantage of the complementary benefits from the CNN and GAT architectures.

On the other hand, results of subsets where our approach is not competitive also bear some relevant insights. First, further research is needed in the expression subset, where our performance is not as good as the rest. This is due to the fact that the 3D facial model used to initialize the cascade is rigid (see Fig. 3). Second, seemingly, in the pose subset, we are not the top performers. However, as we can see in Fig. 3, faces with extreme poses are not well annotated and self-occlusions are not marked. So, the evaluation on this subset of WFLW is

[1] Use RetinaFace detections.

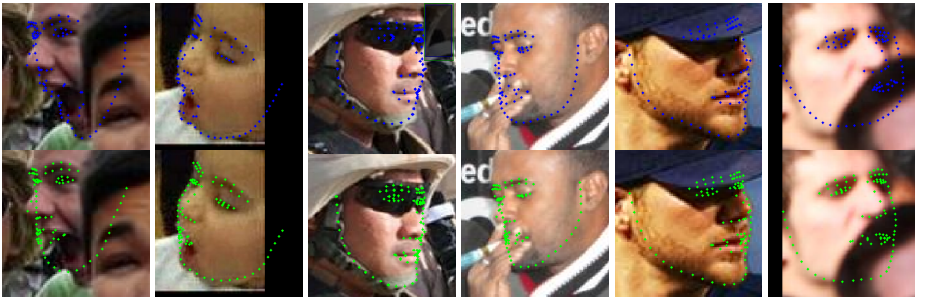

Figure 3: WFLW results on expressions (first 2 cols.) and pose examples (last 4 cols.). Shown in blue the ground truth and in green estimated landmarks.

Method	$NME_{box}(\%)(\downarrow)$ All	Frontal	Half-Prof.	Profile	$AUC^{7}_{box}(\%)(\uparrow)$ All	Frontal	Half-Prof.	Profile
DU-Net	1.99	1.89	2.50	1.92	71.80	73.25	64.78	72.79
LUVLI [13]	1.61	1.74	1.79	1.25	77.08	75.33	74.69	82.10
SPIGA (Ours)	**1.51**	**1.62**	**1.68**	**1.19**	**78.47**	**76.96**	**75.64**	**83.00**

Table 3: Evaluation of landmark detection on MERL-RAV.

questionable.

MERL-RAV is one of the newest datasets, created to evaluate 2D facial alignment in-the-wild. It improves landmark annotations at half-profile and profile images by labeling the self-occlusion of landmarks. Hence, this dataset allows to correctly measure the performance of landmark detectors on samples with extreme poses. As we can see in Table 3, in terms of NME_{box}, our model is 6% better than LUVLI's [13] baseline, performing the best in all pose subsets.

Finally, to verify the generalization and performance against occlusions, we conduct a cross-dataset experiment training with the 300W public split and testing with COFW-68 and 300W private. Results are summarized in Table 4. They prove the importance of the graph attention mechanism, which dynamically weighs landmark relationships according to the local image appearance and relative position, versus a learned static relationship approach, such as SLD [16], ($NME_{int-ocul}$ of 3.93 vs 4.22 in COFW-68). Further, SPIGA trained on the 300W public dataset beats LUVLI [13] (NME_{box} of 2.52 vs 2.75 in COFW-68) with a backbone that has half the number of HG modules. It also obtains comparable results to a recent transformer-based method trained from scratch, DTLD-s [15]. It is marginally better than DTLD-s in 300W private and worse in COFW-68. These results prove that a general architecture using GATs can complement and enhance CNN-based models, reaching better results in situations where ambiguity or noise is contaminating the local landmark appearance, where preserving structural landmarks consistency contributes to the final solution.

3.4 Ablation Study

We conduct our ablation study on WFLW to understand how SPIGA components impact specific subset metrics. Table 5 shows that the addition of the cascade shape regressor outperforms the bare MTN backbone (using `SoftArgMax`). Our new relative positional encoding is better than stacking the vector $\mathbf{q}_t^l$ with the visual features, and much better than

[2]Result comes from a personal communication with authors of [37], 2.09 mistakenly in the paper.

	NME_{box} (%)(↓)		AUC^{7}_{box} (%)(↑)		$NME_{int-ocul}$ (%)(↓)
	300W priv.	COFW-68	300W priv.	COFW-68	COFW-68
HRNetV2-W18 [25]	-	-	-	-	5.06
HG×1+SAAT [34]	-	-	-	-	4.61
LUVLI(8) [18]	2.24	2.75	68.3	60.8	-
GlomFace [31]	-	2.69 [2]	-	-	4.21
SLD [14]	-	-	-	-	4.22
SDFL [20]	-	-	-	-	4.18
SPLT [33]	-	-	-	-	4.10
DTLD-s [18]	2.05	**2.47**	70.9	**65.0**	-
SPIGA(4) (ours)	**2.03**	2.52	**71.0**	64.1	**3.93**

Table 4: Landmark detection results on 300W private and COFW-68. In $(\cdot)$ we show the number of HG modules.

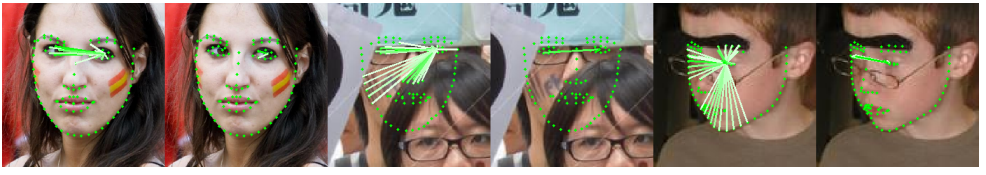

Figure 4: Left pupil attention mechanism at first and last layer, respectively, of the first regressor step.

using no positional information. The estimation of an attention per layer with the GAT improves with respect to use of a common attention matrix (GCN). An extended view of the effect of the learned adjacency matrix is shown in Fig. 4. Occlusion images show how the attention mechanism relies on visible landmarks regardless of the layer. The regressor "looks" at distant and unoccluded landmarks at the first GAT layer and then at closer ones in the last layers. The contribution of the proposed coarse-to-fine scheme w.r.t. a constant size window ($w = 8$) or a single pixel window ($w = 1$) is also clear in Table 5. The improvement provided by SPIGA can be seen across all metrics. However, it is more prominent with the hard cases, as demonstrated by the results for the subsets Makeup, Occlusion, and Blur, and the NPE_{90} of the full set.

In each row of Table 6, we display respectively the performance of three SPIGA models configured with one, two and three steps cascade. In each column, we show the NME obtained at each step. The final NME is reduced gradually as we increase the number of steps. Further, shorter cascades tend to have a better NME at the first step (4.17 vs 4.22). However, given also the larger FR they achieve (2.60 vs 2.44), we can conclude that longer cascades focus their first steps on improving their robustness.

Changed from SPIGA model:		Full		Make-up		Occlusion		Blur	
Changed	From → To	NME	NPE_{90}	NME	NPE_{90}	NME	NPE_{90}	NME	NPE_{90}
Shape model	SPIGA → MTN backbone	4.13	6.93	4.06	7.43	5.10	8.58	4.81	7.70
Positional encoding	SPIGA → w/o pos. encod.	4.17	7.07	4.01	6.71	5.03	8.33	4.72	7.52
	SPIGA → stacking	4.09	6.87	3.83	6.47	4.97	8.15	4.68	7.37
Attention	GAT → GCN	4.08	6.79	3.84	6.54	4.98	**8.05**	4.68	7.37
Coarse-to-Fine	$w = 16,8,4 \rightarrow w = 1,1,1$	4.12	6.95	3.88	6.76	4.99	8.19	4.71	7.44
	$w = 16,8,4 \rightarrow w = 8,8,8$	4.08	6.84	3.82	6.53	4.98	8.13	4.67	7.43
-	**Best SPIGA model**	**4.06**	**6.76**	**3.81**	**6.32**	**4.95**	8.09	**4.65**	**7.31**

Table 5: Contribution of the SPIGA components to the $NME_{int-ocul}(\downarrow)$ and $NPE_{90}(\downarrow)$ in WFLW.

Method	Step 1			Step 2			Step 3		
	$NME_{int-ocul}$ (↓)	AUC_{10} (↑)	FR_{10} (↓)	$NME_{int-ocul}$ (↓)	AUC_{10} (↑)	FR_{10} (↓)	$NME_{int-ocul}$ (↓)	AUC_{10} (↑)	FR_{10} (↓)
SPIGA(1)	4.17	59.53	2.60	-	-	-	-	-	-
SPIGA(2)	4.17	59.55	2.44	4.07	60.45	2.20	-	-	-
SPIGA(3)	4.22	59.10	2.44	4.08	60.41	2.12	4.06	60.56	2.08

Table 6: SPIGA results for cascades with different number of steps, shown in ().

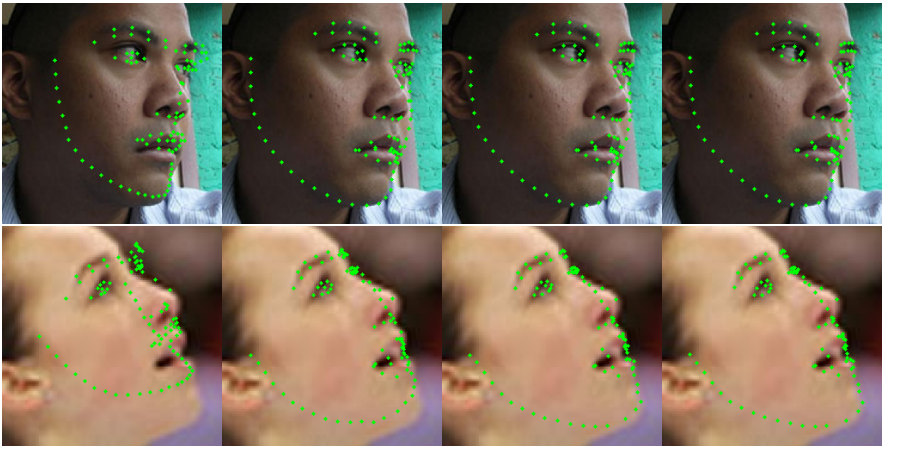

Figure 5: Estimated landmark locations: from 2D projection of the rigid 3D model (left) to the final result after the 3 regressor steps (right).

In Fig. 5 we show the initialization and the landmark locations estimated at each step of the regressor cascade. When the face displays a neutral expression (top row), the initialization is reasonably good and the model converges to a solution within one regression step. Since SPIGA initializes landmarks with a 3D model featuring a neutral expression, when the face displays any other configuration, the initialization is much worse (lower row). However, even in this situation, the model is able to estimate the correct landmark locations in three regression steps.

4 Conclusions

We presented SPIGA, a face landmark regressor that combines a CNN with a cascade of Graph Attention Networks (GATs). The CNN provides the local appearance representation. The GAT regressor is endowed with a positional encoding and attention mechanism that learn the geometrical relationship among landmarks and encourage the model to produce plausible face shapes. It establishes a new SOTA in the WLFW, COFW-68 and MERL-RAV datasets. In our experimentation we verify that the positional encoding is the component that contributes most to the final result and the first steps of the cascade focus on improving the robustness. In addition, at each step, the regressor "looks" at distant and reliable landmarks in the first GAT layer and progressively focuses its attention on closer landmarks in the following ones. These insights from our ablation analysis confirm that SPIGA is learning a global representation and explains why its improvement is most significative in challenging situations involving occlusions, heavy make-up, blur and extreme illumination.

Acknowledgements

The following funding is gratefully acknowledged. Andrés Prados was funded by the Comunidad de Madrid, Ayudantes de Investigación grant PEJ-2019-AI/TIC-15032. The research leading to these results has received funding from RoboCity2030-DIH-CM, Madrid Robotics Digital Innovation Hub, S2018/NMT-4331, funded by "Programas de Actividades I+D en la Comunidad de Madrid" and cofunded by Structural Funds of the EU.

References

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

[2] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *IJCV*, 107(2):177–190, 2014.

[3] Arnaud Dapogny, Matthieu Cord, and Kevin Bailly. Decafa: Deep convolutional cascade for face alignment in the wild. In *ICCV*, pages 6892–6900. IEEE, 2019.

[4] Piotr Dollar, Peter Welinder, and Pietro Perona. Cascaded pose regression. In *CVPR*, pages 1078–1085, 2010.

[5] 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.

[6] ZH. Feng, J. Kittler, M. Awais, and Xiao-Jun Wu. Rectified wing loss for efficient and robust facial landmark localisation with convolutional neural networks. *IJCV*, 128: 2126–2145, 2020.

[7] 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.

[8] Sina Honari, Jason Yosinski, Pascal Vincent, and Christopher J. Pal. Recombinator networks: Learning coarse-to-fine feature aggregation. In *CVPR*, pages 5743–5752, 2016.

[9] Xiehe Huang, Weihong Deng, Haifeng Shen, Xiubao Zhang, and Jieping Ye. Propagationnet: Propagate points to curve to learn structure information. In *CVPR*, June 2020.

[10] 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.

[11] 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.

[12] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *CVPRW*, pages 2034–2043, 2017.

[13] 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.

[14] Xing Lan, Qinghao Hu, and Jian Cheng. Revisting quantization error in face alignment. In *ICCVW*, pages 1521–1530, October 2021.

[15] 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.

[16] 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.

[17] 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.

[18] Olga Moskvyak, Frederic Maire, Feras Dayoub, and Mahsa Baktashmotlagh. Keypoint-aligned embeddings for image retrieval and re-identification. In *WACV*, pages 676–685, January 2021.

[19] 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.

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

[21] A. Santoro, D. Raposo, D. G Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *NeurIPS*, 2017.

[22] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, June 2020.

[23] Ning Sun, Qi Li, Ruizhi Huan, Jixin Liu, and Guang Han. Deep spatial-temporal feature fusion for facial expression recognition in static images. *PRL*, 119:49–61, 2019.

[24] George Trigeorgis, Patrick Snape, Mihalis A. Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *CVPR*, pages 4177–4187, 2016.

[25] 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.

[26] 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.

[27] Roberto Valle, José M. Buenaposada, and Luis Baumela. Multi-task head pose estimation in-the-wild. *IEEE TPAMI*, 43(8):2874–2881, 2021.

[28] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[29] 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.

[30] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *ICCV*, October 2019.

[31] 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.

[32] 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.

[33] Xiang Xu and Ioannis A. Kakadiaris. Joint head pose estimation and face alignment framework using global and local CNN features. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 642–649. IEEE Computer Society, 2017.

[34] Heng Yang, Wenxuan Mou, Yichi Zhang, Ioannis Patras, Hatice Gunes, and Peter Robinson. Face alignment assisted by head pose estimation. In *BMVC*, pages 130.1–130.13, 2015.

[35] Jiangning Zhang, Xianfang Zeng, Mengmeng Wang, Yusu Pan, Liang Liu, Yong Liu, Yu Ding, and Changjie Fan. Freenet: Multi-identity face reenactment. In *CVPR*, pages 5325–5334, 2020.

[36] 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.

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