# Shape Preserving Facial Landmarks with Graph Attention Networks

**Andrés Prados-Torreblanca [1,2], José M. Buenaposada [1], Luis Baumela [2]**

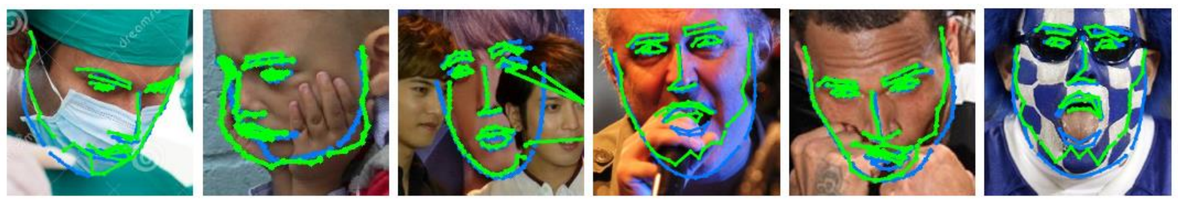
Universidad Rey Juan Carlos [1], Universidad Politécnica de Madrid [2]

BMVC 2022

## Abstract

Facial landmarks estimation is a crucial step for many face analysis problems such as facial expression recognition, face reenactment, etc.

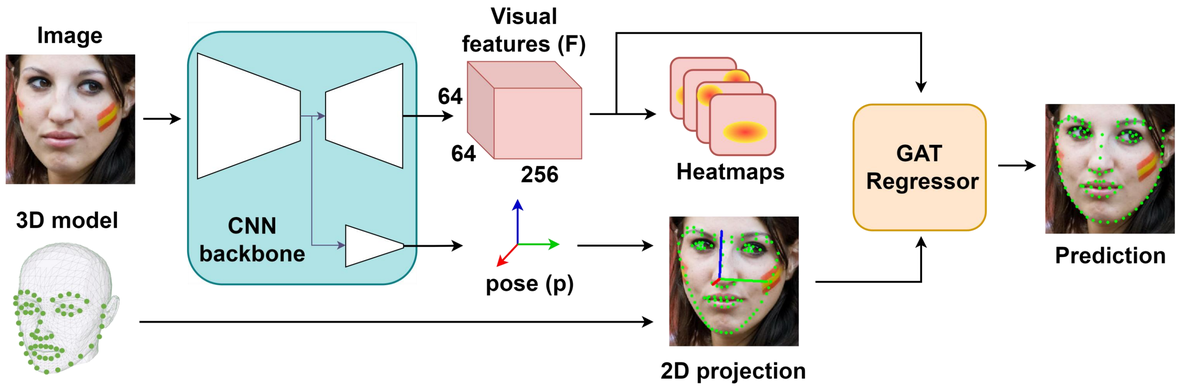
**Problem:** CNN architectures have difficulties to learn simple spatial relationships. In our case, the global representation of the structure of a face.



Therefore, an effective way of representing the local appearance of each landmark and its geometric relationship to the other landmarks is needed.

**Key contributions:** We present SPIGA, a robust method which takes advantage of the complementary benefits from CNN and GNN architectures.

- A GAT cascade with an attention mechanism to weigh the information provided by each landmark according to its reliability.
- A positional encoding to jointly represent relative landmark locations and local appearance.
- A multi-task CNN approach to initialize the location of graph nodes.
- A coarse-to-fine landmark description scheme.




## Experiments

### Pose estimation:

| Method | 300W Angular error (°)(↓) yaw | pitch | roll | mean | WFLW Angular error (°)(↓) yaw | pitch | roll | mean |
|---|---|---|---|---|---|---|---|---|
| ASMNet | 1.62 | 1.80 | 1.24 | 1.55 | 2.97 | 2.93 | 2.21 | 2.70 |
| MNN | - | - | - | 1.56 | - | - | - | 2.08 |
| **SPIGA (Ours)** | **1.41** | **1.70** | **0.77** | **1.29** | **1.78** | **1.86** | **0.93** | **1.52** |

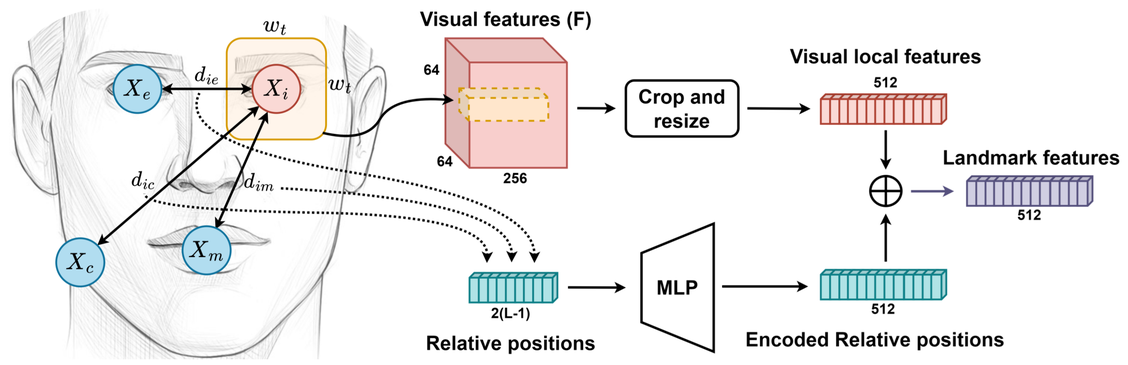
### Landmark estimation results:

| Method | Full | Pose | Expr. | Illum. | Make-up | Occl. | Blur |
|---|---|---|---|---|---|---|---|
| WLFW $NME_{int-ocul}$ (%)(↓) | | | | | | | |
| SLD | 4.21 | 7.36 | 4.49 | 4.12 | 4.05 | 4.98 | 4.82 |
| ADNet | 4.14 | **6.96** | **4.38** | 4.09 | 4.05 | 5.06 | 4.79 |
| SPLT | 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** |
| WFLW $FR_{10}$ (%)(↓) | | | | | | | |
| SLD | 3.04 | 15.95 | 2.86 | 2.72 | **1.46** | 5.29 | 4.01 |
| ADNet | 2.72 | 12.72 | **2.15** | 2.44 | 1.94 | 5.79 | 3.54 |
| SPLT | 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** |

## Qualitative results


## Method

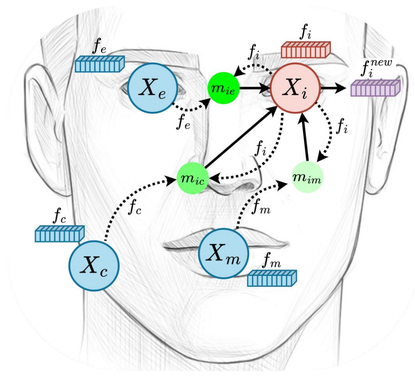
**Relative Positional Encoding:**
Combining visual and geometrical information is crucial to maintain the shape of the face when local appearance alone is not sufficient.




Relative landmark distances explicitly represent the facial shape, providing enhanced geometrical features compared to their absolute locations.

**Graph Attention Networks:**

- We consider the facial shape as a stable and fully-connected weighted graph where nodes are landmarks.



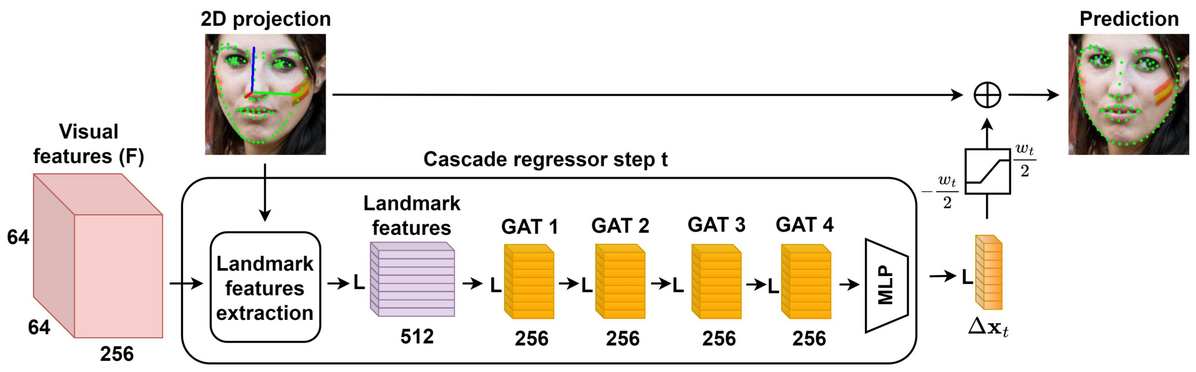

**Node feature update**

$$\mathbf{f}_{i,s+1} = \mathbf{f}_{i,s} + MLP([\mathbf{f}_{i,s} || \mathbf{m}_{i,s}])$$

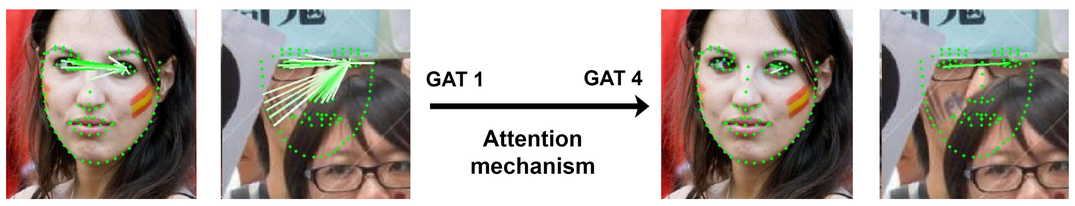
**Attentional message passing**

$$\mathbf{m}_i = \sum_{i \neq j} \alpha(\mathbf{f}_i, \mathbf{f}_j) \Phi_v(\mathbf{f}_j)$$

- To weigh the shared information across nodes, we compute a dynamic adjacency matrix per GAT layer.




- The attention mechanism initially attends to a wide range of landmarks (GAT1) and gradually focus on specific ones (GAT4).




- Occlusion images show how the attention mechanism relies on visible landmarks.

**Coarse-to-fine Cascaded Regressor:** Constraining the output step displacement to the visual cropped region boost training convergence.