# Subpixel Heatmap Regression for Facial Landmark Localization

Adrian Bulat[1]
adrian@adrianbulat.com
Enrique Sanchez[1]
e.lozano@samsung.com
Georgios Tzimiropoulos[1,2]
g.tzimiropoulos@qmul.ac.uk

[1] Samsung AI Center
Cambridge, UK

[2] Queen Mary University London
London, UK

**Abstract**

Deep Learning models based on heatmap regression have revolutionized the task of facial landmark localization with existing models working robustly under large poses, non-uniform illumination and shadows, occlusions and self-occlusions, low resolution and blur. However, despite their wide adoption, heatmap regression approaches suffer from discretization-induced errors related to both the heatmap encoding and decoding process. In this work we show that these errors have a surprisingly large negative impact on facial alignment accuracy. To alleviate this problem, we propose a new approach for the heatmap encoding and decoding process by leveraging the underlying continuous distribution. To take full advantage of the newly proposed encoding-decoding mechanism, we also introduce a Siamese-based training that enforces heatmap consistency across various geometric image transformations. Our approach offers noticeable gains across multiple datasets setting a new state-of-the-art result in facial landmark localization. Code alongside the pretrained models will be made available here.

## 1 Introduction

This paper is on the popular task of localizing landmarks (or keypoints) on the human face, also known as facial landmark localization or face alignment. Current state-of-the-art is represented by fully convolutional networks trained to perform heatmap regression [5, 13, 20, 37, 41, 43]. Such methods can work robustly under large poses, non-uniform illumination and shadows, occlusions and self-occlusions [3, 5, 20, 38] and even very low resolution [6]. However, despite their wide adoption, heatmap-based regression approaches suffer from discretization-induced errors. Although this is in general known, there are very few papers that study this problem [25, 39, 42]. Yet, in this paper, we show that this overlooked problem makes actually has surprisingly negative impact on the accuracy of the model.

In particular, as working in high resolutions is computationally and memory prohibitive, typically, heatmap regression networks make predictions at $\frac{1}{4}$ of the input resolution [5]. Note that the input image may already be a downsampled version of the original facial image. Due to the heatmap construction process that discretizes all values into a grid and

the subsequent estimation process that consists of finding the coordinates of the maximum, large discretization errors are introduced. This in turn causes at least two problems: (a) the encoding process forces the network to learn randomly displaced points and, (b) the inference process of the decoder is done on a discrete grid failing to account for the continuous underlying Gaussian distribution of the heatmap.

To alleviate the above problem, in this paper, we make the following **contributions**:

- We rigorously study and propose a continuous method for heatmap regression, consisting of a simple continuous heatmap encoding and a newly proposed continuous heatmap decoding method, called local-softargmax, that largely solve the quantization errors introduced by the heatmap discretization process.
- We also propose an accompanying Siamese-based training procedure that enforces consistent heatmap predictions across various geometric image transformations.
- By largely alleviating the quantization problem with the proposed solutions, we show that the standard method of [5] sets a new state-of-the-art on multiple datasets, offering significant improvements over prior-work.

## 2 Related work

Most recent efforts on improving the accuracy of face alignment fall into one of the following two categories: network architecture improvements and loss function improvements.

**Network architectural improvements:** The first work to popularize and make use of encoder-decoder models with heatmap-based regression for face alignment was the work of Bulat&Tzimiropoulos [5] where the authors adapted an HourGlass network [27] with 4 stages and the Hierarchical Block of [4] for face alignment. Subsequent works generally preserved the same style of U-Net [32] and Hourglass structures with notable differences in [38, 43, 47] where the authors used ResNets [16] adapted for dense pixel-wise predictions. More specifically, in [47], the authors removed the last fully connected layer and the global pooling operation from a ResNet model and then attempted to recover the lost resolution using a series of convolutions and deconvolutional layers. In [43], Wang *et al*. expanded upon this by introducing a novel structure that connects high-to-low convolution streams in parallel, maintaining the high-resolution representations through the entire model. Building on top of [5], in CU-Net [41] and DU-Net [40] Tang *et al*. combined U-Nets with DenseNet-like [17] architectures connecting the *i*-th U-Net with all previous ones via skip connections.

**Loss function improvements:** The standard loss typically used for heatmap regression is a pixel-wise $\ell_2$ or $\ell_1$ loss [2, 3, 5, 37, 41, 43]. Feng *et al*. [13] argued that more attention should be payed to small and medium range errors during training, introducing the Wing loss that amplifies the impact of the errors within a defined interval by switching from an $\ell_1$ to a modified log-based loss. Improving upon this, in [44], the authors introduced the Adaptive Wing Loss, a loss capable to update its curvature based on the ground truth pixels. The predictions are further aided by the integration of coordinates encoding via CoordConv [24] into the model. In [20], Kumar *et al*. introduced the so-called LUVLi loss that jointly optimizes the location of the keypoints, the uncertainty, and the visibility likelihood. Albeit for human pose estimation, [25] proposes an alternative to heatmap-based regression by introducing a differential soft-argmax function applied globally to the output features. However, the lack of structure induced by a Gaussian prior, hinders their accuracy.

Contrary to the aforementioned works, we attempt to address the quantization-induced error by proposing a simple continuous approach to the heatmap encoding and decoding process. In this direction, [39] proposes an analytic solution to obtain the fractional shift by assuming that the generated heatmap follows a Gaussian distribution and applies this to stabilize facial landmark localization in video. A similar assumption is made by [42] which solves an optimization problem to obtain the subpixel solution. Finally, [25] uses global softargmax. Our method is mostly similar to [25] which we compare with in Section 4.

# 3 Method

## 3.1 Preliminaries

Given a training sample $(\mathbf{X},\mathbf{y})$, with $\mathbf{y} \in \mathbb{R}^{k\times 2}$ denoting the coordinates of the $K$ joints in the corresponding image $\mathbf{X}$, current facial landmark localization methods encode the target ground truth coordinates as a set of $k$ heatmaps with a 2D Gaussian centered at them:

$$\mathcal{G}_{i,j,k}(\mathbf{y}) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}[(i-\tilde{y}_k^{[1]})^2+(j-\tilde{y}_k^{[2]})^2]}, \tag{1}$$

where $y_k^{[1]}$ and $y_k^{[2]}$ are the spatial coordinates of the $k$-th point, and $\tilde{y}_k^{[1]}$ and $\tilde{y}_k^{[2]}$ their scaled, quantized version:

$$(\tilde{y}_k^{[1]},\tilde{y}_k^{[2]}) = (\lfloor \frac{1}{s} y_k^{[1]} \rceil, \lfloor \frac{1}{s} y_k^{[2]} \rceil) \tag{2}$$

where $\lfloor . \rceil$ is the rounding operator and $1/s$ is the scaling factor used to scale the image to a pre-defined resolution. $\sigma$ is the variance, a fixed value which is task and dataset dependent. For a given set of landmarks $\mathbf{y}$, Eq. 1 produces a corresponding heatmap $\mathcal{H} \in \mathbb{R}^{k\times W_{hm}\times H_{hm}}$.

Heatmap-based regression overcomes the lack of a spatial and contextual information of direct coordinate regression. Not only such representations are easier to learn by allowing visually similar parts to produce proportionally high responses instead of predicting a unique value, but they are also more interpretable and semantically meaningful.

## 3.2 Continuous Heatmap Encoding

Despite the advantages of heatmap regression, one key inherent issue with the approach has been overlooked: The heatmap generation process introduces relatively high quantization errors. This is a direct consequence of the trade-offs made during the generation process: since generating the heatmaps predictions at the original image resolution is prohibitive, the localization process involves cropping and re-scaling the facial images such that the final predicted heatmaps are typically at a $64 \times 64$px resolution [5]. As described in Section 3.1, this process re-scales and quantizes the landmark coordinates as $\hat{\mathbf{y}} = \texttt{quantize}(\frac{1}{s}\mathbf{y})$, where `round` or `floor` is the quantization function. However, there is no need to quantize. One can simply create a Gaussian located at:

$$(\tilde{y}_k^{[1]},\tilde{y}_k^{[2]}) = (\frac{1}{s} y_k^{[1]}, \frac{1}{s} y_k^{[2]}), \tag{3}$$

and then sample it over a regular spatial grid. This will completely remove the quantization error introduced previously and will only add some aliasing due to the sampling process.

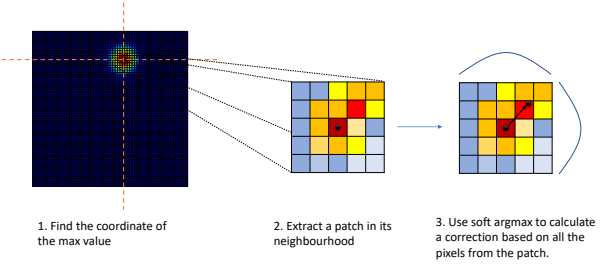



Figure 1: **Proposed heatmap decoding.** Given a predicted heatmap, (1) we find the location of the maximum, (2) and then crop around it a $k \times k$ patch. Finally, (3) we apply a soft-argmax on the patch and retrieve a correction applied to the location estimated at step (1).

### 3.3 Continuous Heatmap Decoding with Local Soft-argmax

Currently, the typical landmark localization process from 2D heatmaps consists of finding the location of the pixel with the highest value [5]. This is typically followed by a heuristic correction with 0.25px toward the location of the second highest neighboring pixel. The goal of this adjustment is to partially compensate for the effect induced by the quantization process: on one side by the heatmap generation process itself (as described in Section 3.2) and on other side, by the coarse nature of the predicted heatmap that uses the maximum value solely as the location of the point. We note that, despite the fact that the ground truth heatmaps are affected by quantization errors, generally, the networks learns to adjust, to some extent its predictions, making the later heuristic correction work well in practice.

Rather than using the above heuristic, we propose to predict the location of the keypoint by analyzing the pixels in its neighbourhood and exploiting the known targeted Gaussian distribution. For a given heatmap $\mathcal{H}_k$, we firstly find the coordinates corresponding to the maximum value $(\hat{y}_k^{[1]}, \hat{y}_k^{[2]}) = \arg\max \mathcal{H}_k$ and then, around this location, we select a small square matrix $h_k$ of size $d \times d$, where $l = \frac{d}{2}$. Then, we predict an offset $(\Delta\hat{y}_k^{[1]}, \Delta\hat{y}_k^{[2]})$ by finding a soft continuous maximum value within the selected matrix, effectively retrieving a correction, using a local soft-argmax:

$$(\Delta\hat{y}_k^{[1]}, \Delta\hat{y}_k^{[2]}) = \sum_{m,n} \texttt{softmax}(\tau h_k)_{m,n}(m,n), \tag{4}$$

where $\tau$ is the temperature that controls the resulting probability map, and $(m,n)$ are the indices that iterate over the pixel coordinates of the heatmap $h_k$. $\texttt{softmax}$ is defined as:

$$\texttt{softmax}(h)_{m,n} = \frac{e^{h_{m,n}}}{\sum_{m',n'} e^{h_{m',n'}}} \tag{5}$$

The final prediction is then obtained as: $(\hat{y}_k^{[1]} + \Delta\hat{y}_k^{[1]} - l, \hat{y}_k^{[2]} + \Delta\hat{y}_k^{[2]} - l)$. The 3 step process is illustrated in Fig. 1.

### 3.4 Siamese consistency training

Largely, the face alignment training procedure has remained unchanged since the very first deep learning methods of [5, 54]. Herein, we propose to deviate from this paradigm adopting a Siamese-based training, where two different random augmentations of the same image are

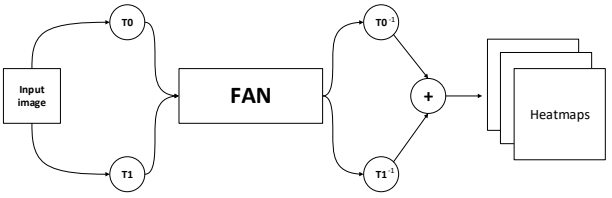



Figure 2: **Siamese transformation-invariant training.** T0 and T1 are two randomly sampled data augmentation transformations applied on the input image. After passing the augmented images through the network a set of heatmaps are produced. Finally, the transformations are reversed and the two outputs merged.

passed through the network, producing in the process a set of heatmaps. We then revert the transformation of each of these heatmaps and combine them via element-wise summation.

The advantages of this training process are twofold: Firstly, convolutional networks are not invariant under arbitrary affine transformations, and, as such, relatively small variances in the input space can result in large differences in the output. Therefore, by optimizing jointly and combining the two predictions we can improve the consistency of the predictions.

Secondly, while previously the 2D Gaussians were always centered around an integer pixel location due to the quantization of the coordinates via rounding, the newly proposed heatmap generation can have the center in-between (*i.e.* on a sub-pixel). As such, to avoid small sub-pixel inconsistencies and misalignment introduced by the data augmentation process we adopt the above-mentioned Siamese based training. Our approach, depicted in Fig. 2, defines the output heatmaps $\tilde{\mathcal{H}}$ as:

$$\tilde{\mathcal{H}} = T_0^{-1}(\Phi(T_0(\mathbf{X}_i),\theta)) + T_1^{-1}(\Phi(T_1(\mathbf{X}_i)),\theta), \tag{6}$$

where $\Phi$ is the network for heatmap regression with parameters $\theta$. $T_0$ and $T_1$ are two random transformations applied on the input image $\mathbf{X}_i$ and, $T_0^{-1}$ and $T_1^{-1}$ denote their inverse.

# 4 Ablation studies

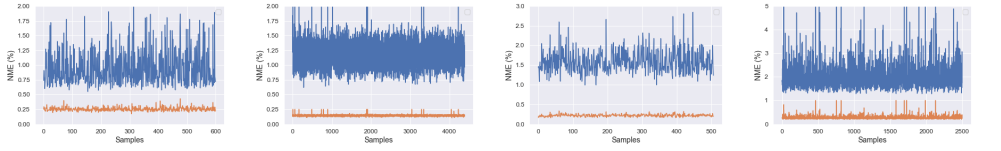
## 4.1 Comparison with other landmarks localization losses

Beyond comparisons with recently proposed methods for face alignment in Section 6 (e.g. [13, 20, 44]), herein we compare our approach against a few additional baselines.

**Heatmap prediction with coordinate correction:** In DeepCut [29], for human pose estimation, the authors propose to add a coordinate refinement layer that predicts a $(\Delta\hat{y}_k^{[1]}, \Delta\hat{y}_k^{[2]})$ displacement that is then added to the integer predictions generated by the heatmaps. To implement this, we added a global pooling operation followed by a fully connected layer and then trained it jointly using an $\ell_2$ loss. We attempted 2 different variants: one where the $(\Delta\hat{y}_k^{[1]}, \Delta\hat{y}_k^{[2]})$ is constructed by measuring the heatmap encoding errors and the other is dynamically constructed at runtime by measuring the error between

| Method | $\text{NME}_{box}$ |
|---|---|
| $\ell_2$ heatmap regression | 2.32 |
| coord-correction (static gt) | 2.27 |
| coord-correction (dynamic gt) | 2.30 |
| Global soft-argmax | 3.19 |
| Local soft-argmax (Ours) | 2.04 |

Table 1: Comparison between various losses baselines on 300W test set.




(a) 300W (b) AFLW (c) COFW (d) WFLW

Figure 3: NME after encoding and then decoding of the ground truth heatmaps for various datasets using our proposed approach (orange) and the standard one [5] (blue). Notice that our approach significantly reduces the error rate across all samples from the datasets.

the heatmap prediction and the ground truth. As Table 1 shows, these learned corrections offer minimal improvements on top of the standard heatmap regression loss and are noticeably worse than the accuracy scored by the proposed method. This shows that predicting sub-pixel errors using a second branch is less effective than constructing better heatmaps from the first place.

**Global soft-argmax:** In [25], the authors propose to to predict the locations of the points of interest on the human body by estimating their position using a global soft-argmax as a differentiable alternative to taking the argmax. From a first glance this is akin to the idea proposed in this work: local soft-argmax. However, applying soft-argmax globally leads to semantically unstructured outputs [25] that hurt the performance. Even adding a Gaussian prior is insufficient for achieving high accuracy on face alignment. As the results from Table 1 conclusively show, our simple improvement, namely the proposed local soft-argmax is the key idea for obtaining highly accurate results.

## 4.2 Effect of method's components

Herein, we explore the impact of each our method's component on the overall performance of the network. As the results from Table 2 show, starting from the baseline introduced in [5], the addition of the proposed heatmap encoding and decoding process significantly improves the accuracy. If we analyze this result in tandem with Fig. 3 it becomes apparent what is the source of these gains: In particular, Fig. 3 shows the heatmap encoding and decoding process of the baseline method [5] as well as of our method using directly the ground truth landmarks (i.e. these are not network's predictions). As shown in Fig. 3, simply encoding and decoding the heatmaps corresponding to the ground truth alone induces high NME for [5]. While the training procedure is able to compensate this, these inaccuracies representations hinder the learning process. Furthermore, due to the sub-pixel errors introduced, the performance in the high accuracy regime of the cumulative error curve degrades.

| Method | $NME_{ic}$ (%) |
|---|---|
| Baseline [5] | 4.20 |
| + proposed hm | 3.90 |
| + proposed hm (w/o 3.3) | 4.00 |
| + siamese training | 3.72 |

Table 2: Effect of the proposed components on the WFLW dataset.

The rest of the gains are achieved by switching to the proposed Siamese training that reduces the discrepancies between multiple views of the same image while also reducing potential sub-pixel displacements that may occur between the image and the heatmaps.

## 4.3 Local window size

In this section, we analyze the relation between the local soft-argmax window size and the model's accuracy. As the results from Table 3 show, the optimal window has a size of

$5 \times 5$px, which corresponds to the size of the generated gaussian (i.e., most of the non-zero values will be contained within this window). Furthermore, as the window size increases the amount of noise and background pixels also increases and hence the accuracy decreases. The same value is used across all datasets. Note, that explicitly using the local window loss during training doesn't improve the performance further which suggest that the pixel-wise loss alone is sufficient, if the encoding process is accurate.

# 5 Experimental setup

**Datasets:** We preformed extensive evaluations to quantify the effectiveness of the proposed method. We trained and/or tested our method on the following datasets: 300W [33] (constructed in [33] using images from LFPW [1], AFW [53], HELEN [21] and iBUG [34]), 300W-LP [54], Menpo [50], COFW-29 [7], COFW-68 [14], AFLW [19], WFLW [45] and 300VW [36]. For a detailed description of each dataset see supplementary material.

| | none | $3 \times 3$ | $5 \times 5$ | $7 \times 7$ |
|---|---|---|---|---|
| $\text{NME}_{box}$ | 2.21 | 2.06 | 2.04 | 2.07 |

Table 3: Effect of window size on the 300W test set.

**Metrics:** Depending on the evaluation protocol of each dataset we used one or more of the following metrics:
**Normalized Mean Error (NME)** that measures the point-to-point normalized Euclidean distance. Depending on the testing protocol, the NME *type* will vary. In this paper, we distinguish between the following types: $d_{ic}$ – computed as the inter-occular distance [33], $d_{box}$ – computed as the geometric mean of the ground truth bounding box [6] $d = \sqrt{(w_{bbox} \cdot h_{bbox})}$, and finally $d_{diag}$ – defined as the diagonal of the bounding box.
**Area Under the Curve(AUC):** The AUC is computed by measuring the area under the curve up to a given user defined cut-off threshold of the cumulative error curve.
**Failure Rate (FR):** The failure rate is defined as the percentage of images the NME of which is bigger than a given (large) threshold.

## 5.1 Training details

For training the models used throughout this paper we largely followed the common best practices from literature. Mainly, during training we applied the following augmentation techniques: Random rotation (between $\pm 30^o$), image flipping and color(0.6, 1.4) and scale jittering (between 0.85 and 1.15). The models where trained for 50 epochs using a step scheduler that dropped the learning rate at epoch 25 and 40 starting from a starting learning rate of 0.0001. Finally, we used Adam [18] for optimization. The predicted heatmaps were at a resolution of $64 \times 64$px, i.e. $4\times$ smaller than the input images which were resized to $256 \times 256$ pixels with the face size being approximately equal to $220 \times 220$px. The network was optimized using an $\ell_2$ pixel-wise loss. For the heatmap decoding process, the temperature of the soft-argmax $\tau$ was set to 10 for all datasets, however slightly higher values perform similarly. Values that are too small or high would ignore and respectively overly emphasise the pixels found around the coordinates of the max. All the experiments were implemented using PyTorch [28] and Kornia [31].
**Network architecture:** All models trained throughout this work, unless otherwise specified, follow a 2-stack Hourglass based architecture with a width of 256 channels, operating at a

resolution of $256 \times 256$px as introduced in [5]. Inside the hourglass, the features are rescaled down-to $4 \times 4$px and then upsampled back, with skip connection linking features found at the same resolution. The network is constructed used the building block from [4] as in [5]. For more details regarding the network structure see [5, 27].

# 6 Comparison against state-of-the-art

Herein, we compare against the current state-of-the-art face alignment methods across a plethora of datasets. Throughout this section the best result is marked in table with bold and red while the second best with bold and blue color. The important finding of this section is by means of two simple improvements: (a) improving the heatmap encoding and decoding process and, (b) including the Siamese training, we managed to obtain results which are significantly better than all recent prior work, setting in this way a new state-of-the-art.

**Comparison on WFLW:** On WFLW, and following their evaluation protocol, we report results in terms of $\text{NME}_{ic}$, $\text{AUC}_{ic}^{10}$ and $\text{FR}_{ic}^{10}$. As the results from Table 4a show, our method improves the previous best results of [20] by more than 0.5% for $\text{NME}_{ic}$ and 5% in terms of $\text{AUC}_{ic}^{10}$ almost halving the error rate. This shows that our method offers improvements in the high accuracy regime while also reducing the overall failure ratio for difficult images.

**Comparison on AFLW:** Following [20], we report results in terms of $\text{NME}_{diag}$, $\text{NME}_{box}$ and $\text{AUC}_{box}^{7}$. As the results from Table 6 show, we improve across all metrics on top of the current best result even on this nearly saturated dataset.

| Method | $\text{NME}_{ic}(\%)$ | $\text{AUC}_{ic}^{10}$ | $\text{FR}_{ic}^{10}$ (%) |
|---|---|---|---|
| Wing [13] | 5.11 | 0.554 | 6.00 |
| MHHN [42] | 4.77 | - | |
| DeCaFa [9] | 4.62 | 0.563 | 4.84 |
| AVS [30] | 4.39 | **0.591** | 4.08 |
| AWing [42] | 4.36 | 0.572 | **2.84** |
| LUVLi [20] | 4.37 | 0.577 | 3.12 |
| GCN [23] | **4.21** | 0.589 | 3.04 |
| Ours | **3.72** | **0.631** | **1.55** |

(a) Comparison against the state-of-the-art on WFLW in terms of $\text{NME}_{inter-ocular}$, $\text{AUC}_{ic}^{10}$ and $\text{FR}_{ic}^{10}$.

| | Common | Challenge | Full |
|---|---|---|---|
| Teacher [10] | 2.91 | 5.91 | 3.49 |
| DU-Net [41] | 2.97 | 5.53 | 3.47 |
| DeCaFa [9] | 2.93 | 5.26 | 3.39 |
| HR-Net [38] | 2.87 | 5.15 | 3.32 |
| HG-HSLE [55] | 2.85 | 5.03 | 3.28 |
| Awing [42] | **2.72** | **4.52** | **3.07** |
| LUVLi [20] | 2.76 | 5.16 | 3.23 |
| Ours | **2.61** | **4.13** | **2.94** |

(b) Comparison against state-of-the-art on the 300W Common, Challenge and Full datasets (*i.e.* Split II) in terms of $\text{NME}_{inter-occular}$

Table 4: Results on WFLW (a) and 300W (b) datasets.

**Comparison on 300W:** Following the protocol described in [33] and [5], we report results in terms of $\text{NME}_{inter-occular}$ for *Split I* and of $\text{AUC}_{box}^{7}$ and $\text{NME}_{box}$ for *split II*. Note that due to the overlap between the splits we train two separate models, one on the data from the first split and another on the data from the other split evaluating the models accordingly. Following [5, 20] the model evaluated on the test set was pretrained on 300W-

| Method | $\text{NME}_{ic}(\%)$ | $\text{FR}_{ic}^{10}$ (%) |
|---|---|---|
| Wing [13] | 5.07 | 3.16 |
| LAB (w/B) [46] | 3.92 | 0.39 |
| HR-Net [38] | **3.45** | **0.19** |
| Ours | **3.02** | **0.0** |

Table 5: Comparison on COFW-29. Results for other methods taken from [38].

| Method | $NME_{diag}$ | | $NME_{box}$ | $AUC^{7}_{box}$ |
|---|---|---|---|---|
| | Full | Frontal | Full | Full |
| SAN [10] | 1.91 | 1.85 | 4.04 | 54.0 |
| DSNR [26] | 1.85 | 1.62 | - | - |
| LAB (w/o B) [46] | 1.85 | 1.62 | - | - |
| HR-Net [33] | 1.57 | 1.46 | - | - |
| Wing [13] | - | - | 3.56 | 53.5 |
| KDN [8] | - | - | 2.80 | 60.3 |
| LUVLi [20] | 1.39 | **1.19** | **2.28** | **68.0** |
| MHHN [22] | **1.38** | **1.19** | - | - |
| Ours | **1.31** | **1.12** | **2.14** | **70.0** |

Table 6: Comparison against the state-of-the-art on the AFLW-19 dataset.

| Method | $NME_{box}$ | | | $AUC^{7}_{box}$ | | |
|---|---|---|---|---|---|---|
| | 300-W | Menpo | COFW-68 | 300-W | Menpo | COFW-68 |
| SAN [10] | 2.86 | 2.95 | 3.50 | 59.7 | 61.9 | 51.9 |
| FAN [6] | 2.32 | 2.16 | 2.95 | 66.5 | 69.0 | 57.5 |
| Softlabel [8] | 2.32 | 2.27 | 2.92 | 66.6 | 67.4 | 57.9 |
| KDN [8] | 2.21 | **2.01** | 2.73 | 68.3 | 71.1 | 60.1 |
| LUVLi [20] | **2.10** | 2.04 | **2.57** | **70.2** | **71.9** | **63.4** |
| Ours | **2.04** | **1.95** | **2.47** | **71.1** | **73.0** | **64.9** |

Table 7: Comparison against the state-of-the-art on the 300W Test (*i.e.* Split I), Menpo 2D Frontal and COFW-68 datasets in terms of $NME_{box}$ and $AUC^{7}_{box}$.

LP dataset. As the results from Table 4b show, our approach offers consistent improvements across both subsets (*i.e.* *Common* and *Challenge*), with particularly higher gains on the later. Similar results can be observer in Table 7 for *Split II*.

**Comparison on COFW:** On the COFW dataset we evaluate on both the 29-point (see Table 5) and 68-point configuration (see Table 7) in terms of $NME_{ic}(\%)$ and $FR^{10}_{ic}$ for the 29-point configuration and $NME_{box}$, $AUC^{7}_{box}$ for the other one. As the results from Tables 7 and 5 show, our method sets a new state-of-the-art, reducing the failure rate to 0.0.

**Comparison on Menpo:** Following [20] we evaluate on the frontal sub-set of the Menpo dataset. As Table 7 shows, our method sets a new state-of-the-art result.

**Comparison on 300VW:** Unlike the previous datasets that focus on face alignment for static images, 300VW is a video face tracking dataset. Following [36], we report results in terms of $AUC_{ic}$@0.08 on the most challenging partition of the test set (C). As the results from Table 8 show, despite not exploiting any temporal information and running our method on a frame-by-frame basis, we set a new state-of-the-art, outperforming previous tracking methods trained such as [35] and [15]. Similar results can be observed when evaluating on all 68 points in Table 9.

| Method | **Ours** | DGM [15] | CPM+SRB+PAM [12] | iCCR [35] | [49] | [48] |
|---|---|---|---|---|---|---|
| $AUC_{ic}$@0.08 | **60.10** | 59.38 | **59.39** | 51.41 | 49.96 | 48.65 |

Table 8: Comparison against the state-of-the-art on the 300-VW dataset – category C, in terms of $AUC_{ic}$@0.08 evaluated on the 49 inner points.

| Method | **Ours** | FHR+STA [39] | TSTN [23] | TCDCN [51] | CFSS [52] |
|---|---|---|---|---|---|
| $NME_{ic}$ | **5.84** | **5.98** | 12.80 | 15.0 | 13.70 |

Table 9: Comparison against the state-of-the-art on the 300-VW dataset – category C (i.e., scenario 3), in terms of $NME_{ic}$ evaluated on all 68 points. Results for other methods taken from [39].

Figure 4: Qualitative results. Landmarks shown in white are produced by our method, while the ones in red by the state-of-the-art approach of [5]. Thanks to the proposed heatmap encoding and decoding, our method is able to provide much more accurate results. Best viewed zoomed in, in electronic format.

Figure 5: Examples of failure cases. Most of the failure cases include combinations of low resolution images with extreme poses (1st and 4th image), perspective distortions (5th image) or overlapping faces (3rd image).

## 7 Conclusions

We presented simple yet effective improvements to standard methods for face alignment which are shown to dramatically increase the accuracy on all benchmarks considered without introducing sophisticated changes to existing architectures and loss functions. The proposed improvements concern a fairly unexplored topic in face alignment that of the heatmap encoding and decoding process. We showed that the proposed continuous heatmap regression provides a significantly improved approach for the encoding/decoding process. Moreover, we showed that further improvements can be obtained by considering a simple Siamese training procedure that enforces output spatial consistency of geometrically transformed images. We hope that these improvements will be incorporated in future research while it is not unlikely that many existing methods will also benefit by them.

# References


[1] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *TPAMI*, 2013.

[2] Adrian Bulat and Georgios Tzimiropoulos. Convolutional aggregation of local evidence for large pose face alignment. In *BMVC*, 2016.

[3] Adrian Bulat and Georgios Tzimiropoulos. Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In *ECCV*, 2016.

[4] Adrian Bulat and Georgios Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. *arXiv*, 2017.

[5] 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 *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[6] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2018.

[7] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *ICCV*, 2013.

[8] Lisha Chen, Hui Su, and Qiang Ji. Face alignment with kernel density deep neural network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6992–7002, 2019.

[9] Arnaud Dapogny, Kevin Bailly, and Matthieu Cord. Decafa: deep convolutional cascade for face alignment in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6893–6901, 2019.

[10] Xuanyi Dong and Yi Yang. Teacher supervises students how to learn from partially labeled images for facial landmark detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 783–792, 2019.

[11] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–388, 2018.

[12] Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 360–368, 2018.

[13] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2235–2245, 2018.

[14] Golnaz Ghiasi and Charless C Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. In *CVPR*, 2014.

[15] Muhammad Haris Khan, John McDonagh, and Georgios Tzimiropoulos. Synergy between face alignment and tracking via discriminative global consensus optimization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3791–3799, 2017.

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

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] Martin Köstinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCV-W*, 2011.

[20] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8236–8246, 2020.

[21] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *ECCV*, 2012.

[22] Weijian Li, Yuhang Lu, Kang Zheng, Haofu Liao, Chihung Lin, Jiebo Luo, Chi-Tung Cheng, Jing Xiao, Le Lu, Chang-Fu Kuo, et al. Structured landmark detection via topology-adapting deep graph learning. *arXiv preprint arXiv:2004.08190*, 2020.

[23] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Two-stream transformer networks for video-based face alignment. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2546–2554, 2017.

[24] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *arXiv preprint arXiv:1807.03247*, 2018.

[25] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5137–5146, 2018.

[26] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. Direct shape regression networks for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5040–5049, 2018.

[27] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[29] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016.

[30] Shengju Qian, Keqiang Sun, Wayne Wu, Chen Qian, and Jiaya Jia. Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10153–10163, 2019.

[31] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3674–3683, 2020.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[33] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *CVPR*, 2013.

[34] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. A semi-automatic methodology for facial landmark annotation. In *CVPR*, 2013.

[35] Enrique Sánchez-Lozano, Georgios Tzimiropoulos, Brais Martinez, Fernando De la Torre, and Michel Valstar. A functional regression approach to facial landmark tracking. *IEEE transactions on pattern analysis and machine intelligence*, 40(9):2037–2050, 2017.

[36] 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 *ICCVW*, 2015.

[37] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[38] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019.

[39] Ying Tai, Yicong Liang, Xiaoming Liu, Lei Duan, Jilin Li, Chengjie Wang, Feiyue Huang, and Yu Chen. Towards highly accurate and stable face alignment for high-resolution videos. In *AAAI*, 2019.

[40] Zhiqiang Tang, Xi Peng, Shijie Geng, Lingfei Wu, Shaoting Zhang, and Dimitris Metaxas. Quantized densely connected u-nets for efficient landmark localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 339–354, 2018.

[41] Zhiqiang Tang, Xi Peng, Kang Li, and Dimitris N Metaxas. Towards efficient u-nets: A coupled and quantized approach. *IEEE transactions on pattern analysis and machine intelligence*, 42(8):2038–2050, 2019.

[42] Jun Wan, Zhihui Lai, Jun Liu, Jie Zhou, and Can Gao. Robust face alignment by multi-order high-precision hourglass network. *IEEE Transactions on Image Processing*, 30: 121–133, 2020.

[43] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[44] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6971–6981, 2019.

[45] 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.

[46] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2138, 2018.

[47] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[48] Shengtao Xiao, Shuicheng Yan, and Ashraf A Kassim. Facial landmark detection via progressive initialization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 33–40, 2015.

[49] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial parts responses to face detection: A deep learning approach. In *ICCV*, 2015.

[50] Stefanos Zafeiriou, George Trigeorgis, Grigorios Chrysos, Jiankang Deng, and Jie Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 170–179, 2017.

[51] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):918–930, 2015.

[52] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, 2015.

[53] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*. IEEE, 2012.

[54] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016.

[55] Xu Zou, Sheng Zhong, Luxin Yan, Xiangyun Zhao, Jiahuan Zhou, and Ying Wu. Learning robust facial landmark detection via hierarchical structured ensemble. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 141–150, 2019.