Attention-Driven Cropping for Very High Resolution Facial Landmark Detection

Prashanth Chandran[1,2], Derek Bradley[2], Markus Gross[1,2], and Thabo Beeler[2]

[1]Department of Computer Science, ETH Zurich
[2]DisneyResearch|Studios, Zurich
chandrap@inf.ethz.ch, derek.bradley@disneyresearch.com, grossm@inf.ethz.ch,
thabo.beeler@gmail.com

Abstract

Facial landmark detection is a fundamental task for many consumer and high-end applications and is almost entirely solved by machine learning methods today. Existing datasets used to train such algorithms are primarily made up of only low resolution images, and current algorithms are limited to inputs of comparable quality and resolution as the training dataset. On the other hand, high resolution imagery is becoming increasingly more common as consumer cameras improve in quality every year. Therefore, there is need for algorithms that can leverage the rich information available in high resolution imagery. Naïvely attempting to reuse existing network architectures on high resolution imagery is prohibitive due to memory bottlenecks on GPUs. The only current solution is to downsample the images, sacrificing resolution and quality. Building on top of recent progress in attention-based networks, we present a novel, fully convolutional regional architecture that is specially designed for predicting landmarks on very high resolution facial images without downsampling. We demonstrate the flexibility of our architecture by training the proposed model with images of resolutions ranging from 256 x 256 to 4K. In addition to being the first method for facial landmark detection on high resolution images, our approach achieves superior performance over traditional (holistic) state-of-the-art architectures across ALL resolutions, leading to a general-purpose, extremely flexible, high quality landmark detector.

1. Introduction

Landmark detection is one of the classical machine learning tasks in computer vision, nowadays almost entirely solved via deep neural networks. While these network based detectors provide robust detections, their accuracy directly depends on the image resolution they operate on. While even low-end cameras can capture high resolution imagery nowadays, concurrent GPUs are restricted to operate on low resolution imagery due to limited memory. As a consequence, deep learning algorithms are forced to predict landmarks on imagery that may be several orders of magnitude lower in resolution than what would be available, which naturally amplifies prediction inaccuracies.

When observing how human annotators label images, one might realize that they do so at multiple scales. In the context of facial landmarks, they typically annotate the coarse features, such as for example the jawline, at a low resolution where they have the full context of the face but then zoom into specific areas, such as an eye region, to annotate more accurately. Inspired by this behaviour we propose an end-to-end attention-driven architecture that allows to train deep networks on higher resolution images by automatically defining and focusing on regions of interest instead of considering the face holistically. These regions are identified on a low resolution image proxy and extracted from the original high resolution image. They are then scaled to an appropriate size for the network, which has the benefit of aligning the regions to a canonical crop. The second stage then localizes the landmarks in this frontalized zoom-in, which further reduces variability and increases robustness and accuracy.

Using our novel attention-driven architecture we manage to predict landmarks at resolutions up to 4K on a single GPU, showing significant improvements in prediction accuracy over existing methods which are forced to operate on downsampled imagery. We further demonstrate that the proposed concept applies to a variety of recent network architectures, improving performance for all of them.

Despite the fact that our approach targets high resolution imagery, when applied to traditional lower resolution facial images in-the-wild our method also outperform current state-of-the-art architectures in most cases. Therefore, our proposed method is a general-purpose facial landmark detector with high quality across image scales from low resolution to 4K.

2. Related Work

Before the advent of deep learning, several methods based on cascaded regression [5, 44, 41, 23] were proposed to solve the problem of facial landmark detection. Such methods start with an initial guess of landmarks and refine them using a cascade of machine learning models. In recent years however, deep learning methods have significantly advanced the state of the art in facial landmark detection. For a concise summary, we differentiate and describe the contribution of these methods based on their architecture and their approach to the problem.

In terms of network architecture, existing work can be broadly classified into three categories viz. i) networks that contain a combination of convolutional and fully connected or 'dense' layers ii) fully convolutional networks, and iii) recurrent networks. The former consist of architectures that take an image as input and learn convolutional filters that extract low level and semantic features, which are then flattened and passed onto one or more full connected layers [53, 8, 25, 56, 20, 3, 21, 27, 51, 50, 45, 32, 37, 12, 28, 13, 52, 55, 29]. On the other hand, fully convolutional architectures [39, 26, 30, 4, 46, 47, 42, 35, 38, 9, 54, 36, 10] predict the positions of facial landmarks as heatmaps that encode the probability of a landmark being present at a particular pixel. These architectures have a few advantages, namely (i) translation invariance, (ii) images of different sizes can be used at training and test times, (iii) they provide a guarantee that the predicted landmarks always lie within the domain of the image, and (iv) the representation of landmarks as a heatmap makes the prediction of such networks human interpretable. The final category is recurrent network approaches [40, 1], which are designed to operate on a temporal sequence of images by adding recurrent layers.

Based on their approach towards solving facial landmark detection, the above-mentioned methods can also be classified broadly into i) model based fitting methods, ii) multitask learning, and iii) cascaded or regional models. Model based methods [56, 20, 3, 21, 27] assume an underlying low resolution 3D face model that is parametrically fit to facial images using learned features. Multi-task methods [51, 50, 45, 32, 52] follow the principle of 'auxiliary learning' to jointly infer multiple attributes of the given facial image, such as the person's age, gender etc, in addition to facial landmarks. Region based methods [37, 12, 28, 55] consist of a series of architectures that independently analyze different regions of the face.

Existing facial landmark detectors work well on low resolution imagery. However, when a high resolution image is available at test time, existing algorithms cannot make use of the extra detail present due to several reasons. First, architectures with fully connected layers (including all of the existing region based approaches [37, 12, 28]) can be used only with images of the same size with which they have been trained. This would require the high resolution image to be downsampled to a size compatible with the architecture. Additionally, during their forward pass, networks build large intermediate feature representations before predicting the output, which can prove extremely challenging for training at high resolution. In practice, even resolutions of 512 x 512 have proven difficult to fit on a single GPU.

2.1. Contributions

In this work, building on top of recent advances in deep learning [17, 4, 19], we propose an organic evolution to region-based facial landmark detectors and propose an end-to-end differentiable, fully convolutional, region based facial landmark detector.

- We combine attention driven cropping, introduced by [17] with a differentiable soft-argmax [19] operation to enable the first fully convolutional region based facial landmark detector.
- To the best of our knowledge, our method is the first to demonstrate the ability to both train with and infer facial landmarks on images of resolution up to 4096 x 4096 on a single Nvidia 1080Ti GPU. We show the superiority of our method across multiple resolutions ranging from 256x256, up to 4096x4096 over the naïve upsampling of low resolution landmarks detected with previous state of the art methods.
- Although specifically designed for high resolution imagery, our method generalizes extremely well to unconstrained, in-the-wild settings and often outperforms low resolution state-of-the-art methods (Section 4).

2.2. Available Datasets

300-W [33], 300-VW[34], 300-W-LP [56] are popular datasets for training facial landmark detectors. Similar, but more recent, and larger datasets include [48, 49, 4]. These datasets contain annotations for 68 facial landmarks. While [33, 48] are datasets with only 2D annotations, [56, 34, 4] contain both 2D and 3D annotations. All methods described in Section 2 use one or more of these datasets to train and fine tune their models. Existing datasets consist of low resolution imagery captured in an unconstrained setting as they were intended to be used for "in-the-wild" applications. In contrast, our objective is to train a landmark detector that can make use of the detail present in high resolution facial imagery to precisely localize landmarks. Consequently, we cannot use any of the existing datasets for training. We create a new high quality facial landmark dataset for training and testing our high resolution performance (described in Section 3.3). However, to show the additional benefits of our approach to in-the-wild imagery, we also show experiments on the 300-W [33] and 300-VW [34] datasets.

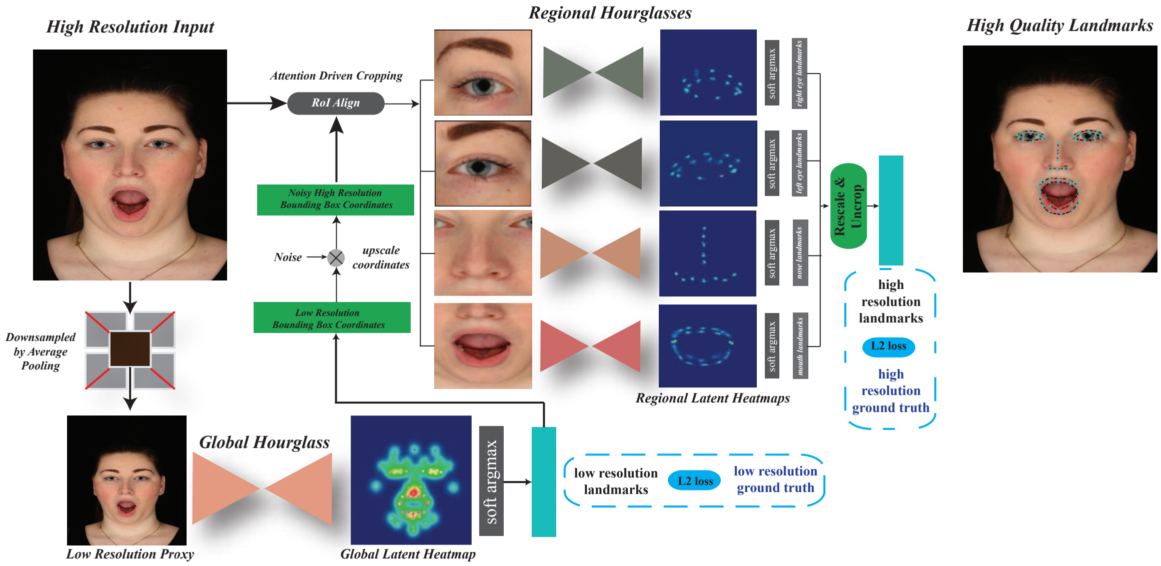

Figure 1. Schematic overview of our attention-driven architecture for facial landmark detection. High resolution input images are downsampled to a low resolution proxy on which a global hourglass network detects low resolution landmarks. Crop regions are automatically determined and the RoIs are re-scaled to the original resolution, where regional hourglass networks detect high resolution landmarks.

3. Methodology

In this section, we present our new architecture for high resolution facial landmark detection, which is depicted in Fig. 1. Inspired by how humans manually annotate landmarks on high resolution images, our model analyzes different regions of the face in isolation, through an attention-driven cropping mechanism. Given an initial high resolution image as the input, a global hourglass network [30] analyzes a corresponding low-resolution proxy of the input image and produces coarse heatmaps of facial landmarks. On these heatmaps, we perform a differentible softargmax operation to extract initial estimates of landmark coordinates. These low resolution landmark coordinates are used to identify regions of interest (ROIs) that correspond to distinct anatomical regions of the face. For each region, a high resolution crop is extracted from the original high resolution input image, and is further analyzed by a region specific hourglass network that predicts landmarks on the crop. The landmarks predicted on the regional crops are restored back to the original image using the ROI information from the global hourglass. Landmarks predicted by both the global and the multiple regional models are supervised with low and high resolution ground truth data, respectively. The proposed architecture is fully differentiable and fully convolutional and can therefore be trained end to end.

3.1. Network Architecture

The input high resolution image is initially downsampled by average pooling to a fixed resolution of 256x256 pixels. The downsampled image is passed through an hourglass network [30] - an architectural choice that is analyzed in Section 4.5 - that outputs heatmaps of the landmark locations at the same scale as the low resolution image. Since the first hourglass predicts landmarks for all regions of the face, we refer to it as the global hourglass. The global hourglass outputs one heatmap for each landmark. Work such as [42, 30, 6, 4], and many others, generate ground truth heatmaps from the training data, from which landmarks are extracted at test time using an argmax operation. Ground truth heatmaps for such methods are generated by applying a spatial gaussian filter on the position of the landmarks. The standard deviation of this gaussian filter σ is manually specified and all landmarks of the face are blurred using the same σ. However, certain landmarks in the training set are localized with higher anisotropic uncertainty due to the underlying feature. In the case of facial landmarks, for example, landmarks like the corners of the eyes and lips are easier to unambiguously identify and therefore to annotate than say the eyelid, where landmarks will have better localization across the edge and higher uncertainty along the edge. Training networks with heatmaps created from isotropic Gaussian kernels is enforcing the assumption that localization of all landmarks is equally (un)certain. Unlike previous methods in facial landmark detection, we choose to represent the output of our convolutional networks as latent heatmaps *without* ground truth supervision. This provides the hourglass network with the flexibility to be more confident about certain landmarks than others and to represent them using anisotropic non-gaussian distributions. Furthermore, we also observe similar improvements in accuracy as reported by Iqbal et.al [19] when using the softargmax over naive heatmap regression. The latent heatmap output by the global hourglass is passed through a channelwise spatial softmax to ensure that each channel is a probability distribution over the landmark's position in the image. Then, we perform a soft-argmax [19] operation on the landmark heatmaps to extract landmark positions as a *batch size x number of landmarks x 2* vector. Since the soft-argmax operation boils down to a weighted average, it is fully differentiable unlike the argmax. Extracting landmark positions this way enables us to train the global hourglass using only the ground truth landmark positions without having to create ground truth heatmaps, while at the same time ensuring that the landmark positions are represented inside the network as a heatmap, and therefore keeping the network fully convolutional.

3.2. Attention-Driven Cropping

In the second stage of our architecture, we use the landmark estimates from the global hourglass to extract regions of interest (RoI) from the original high resolution image. These regions of interest are then individually processed in parallel by a set of region specific hourglass models to refine the position of these landmarks. We refer to these hourglass models that operate on a pre-defined region of the face as regional hourglasses. In this work we train four regional hourglass models, which predict landmarks for the left eye, the right eye, the nose and the mouth regions (please refer to Fig. 2). This approach can be extended to as many ROIs as one would like, but we restrict ourselves to these four regions for the following reason. Outside of these regions, the landmarks that we are interested in belong to the chin, the cheek and the forehead. These regions are typically devoid of salient features, and analyzing these regions locally can in fact be counter-productive due to ambiguities. Such regions are thus better left analyzed globally, at a higher scale by the global hourglass.

For each region of the face, exactly one bounding box is computed using the result of the softargmax. Unlike methods like [15, 14, 17], that generate multiple bounding boxes candidates for each RoI proposal in an 'in the wild' setting, under the assumption that an input image contains only one face, generating a single bounding box per region is reasonable. Each bounding box is represented by 4 co-ordinates corresponding to its top-left and bottom-right corners. Since these bounding box co-ordinates are extracted from the latent heatmap, they are guaranteed to lie within the domain of the downsampled image. Noise from a normal distribution is added to the width and height of the each bounding box independently to make the regional models robust enough to the location of the region inside the bounding box. The noisy bounding boxes are then up-scaled to map them to domain of the original high resolution image. Using the *RoIAlign* operation introduced by [17], we extract crops from the high resolution image in a differentiable manner. The high resolution crops are resized to a fixed size depending on the original resolution of the image. The sizes that we used for the regional crops for different resolutions are shown in table Table 1. The crop sizes were determined based on the original resolution of the image and to maintain a healthy batch size during training. Other crop sizes could also be readily used. The relative scale factors between the noisy high resolution bounding boxes and the resized crop are computed and stored for later restoring the predicted landmarks back to their original resolution. The resized crops are then passed on to the corresponding regional hourglass. Each regional hourglass predicts a latent heatmap of landmark positions similar to the global hourglass. Landmarks defined in the domain of the resized crops are extracted from these regional heatmaps using the softargmax operation as before. These regional landmarks are restored back to the original resolution of the image using the corresponding scale factors computed from before. The rescaled landmarks are then un-cropped using the noisy bounding box co-ordinates to obtain landmarks defined on the high resolution image.

Resolution (pixels)	Crop Size (pixels)	Batch Size
256 x 256	128 x 128	8
512 x 512	128 x 128	8
1024 x 1024	256 x 256	4
2048 x 2048	192 x 192 $*$	4
4096 x 4096	256 x 256 $**$	4

Table 1. Crop sizes used for different image resolutions. $*$ Until a resolution of 2K, we can continue to use the basic hourglass building block. This means that the resolution of the crop increases up to 256 for a 1K input, however this no longer fits onto the GPU when we reach inputs of 2K. Therefore, the size of the crop reduces to 192x192. $**$ For resolutions of 4K, we used the 'light' hourglass variant (Section 3.4), re-enabling crops of higher sizes and as a result, we could use 256x256 crops at 4K.

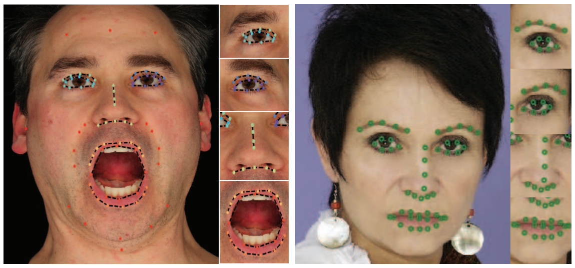

Figure 2. (Left) Our high resolution training data consists of 89 manually-annotated facial landmarks, of which 78 fall within the four attention regions we defined. (Right) 4 attention regions defined on the 300-W dataset corresponding to the two eyes, the nose and the mouth.

Our entire architecture is shown in Fig. 1. Since all operations defined in our architecture are differentiable, the global hourglass and the multiple regional hourglasses can be trained together in an end to end fashion. The final output of our network is a complete set of facial landmark locations for a high resolution image, for which a subset of landmarks (eyes, nose, and mouth) contain high precision locations thanks to our regional refinement modules.

3.3. Training Data

One of the main contributions of our method is that it enables the training of networks with high resolution imagery and sidestep GPU memory bottlenecks via attention-driven cropping. To verify the benefits of our architecture, we require a high resolution dataset of faces with ground truth landmarks. Existing datasets (described in Section 2.2) contain a large number of images in a 'in the wild' setting with 2D annotations but are not of sufficient resolution. To our knowledge, there does not exist an openly available dataset of high resolution facial imagery and landmarks. Therefore we resorted to capturing subjects in a controlled studio setting using the method of [2]. We captured 47 subjects in 4K resolution from 8 cameras performing 24 different facial expressions, and manually annotated 89 facial landmarks on these images. The full set of these 89 landmarks is shown in Fig. 2. Out of the 47 subjects, we randomly sample 24 subjects for training and used the remaining 23 subjects for evaluation. In summary, our training set consisted of a total of 4608 images and our test set consisted of 4416 images. To perform experiments at resolutions of 256 x 256, 512 x 512, 1024 x 1024, 2048 x 2048, and 4096 x 4096, both the training and test sets were appropriately scaled. As seen in Fig. 2, crops are considered only for the regions of the eyes, nose, and the mouth. In our high resolution dataset, out of the 89 annotated landmarks, only 78 fall inside the regional crops. As a result, the global hourglass predicts all 89 landmarks and the regional hourglasses predict a total of 78 landmarks. For 300W, and 300VW, 51 of the 68 landmarks fall under our attention regions. Therefore, while training with 300W and 300VW, our global hourglass would predict 68 landmarks, and the regional hourglasses would predict a total of 51 landmarks.

3.4. Implementation Details

We train the network shown in Fig. 1 by supervising both low and high resolution landmark predictions. The network is trained to minimize the sum of the L2 losses at both resolutions. This additive loss is shown in Eq. 1 where p_n^g and p_n^r correspond to the n^{th} landmark predicted by the global and regional models respectively. gt_n^{lr} and gt_n^{hr} correspond to the n_{th} low and high resolution ground truth respectively. N_{total} and N_{att} correspond to the total, and attention refined landmarks.

$$loss = \frac{1}{N_{total}} \sum_{n=1}^{N_{total}} \|p_n^g - gt_n^{lr}\|^2 + \frac{1}{N_{att}} \sum_{n=1}^{N_{att}} \|p_n^r - gt_n^{hr}\|^2. \tag{1}$$

Though our network is fully convolutional, our use of the soft-argmax enables training with more hand tuned losses like the wingloss [13]. However, since we are interested in analyzing improvement that is obtained by the use of our architecture as opposed to the improvement obtained by using a different loss function, we resorted to using the simple L2 loss in Eq. 1.

We begin by training our architecture at a resolution of 256 x 256. The weights of both the global and regional hourglasses are initialized following [16]. Once training at a resolution of 256 x 256 converges, we begin to train at the next higher resolution of 512 x 512 using the weights from 256 x 256 as an initialization. This initialization is enabled thanks to the fully convolutional nature of our architecture. Likewise, weights are progressively initialized all the way until 4096 x 4096 similar in principle to [22].

An important implementation detail to note is that even with regional models operating on cropped portions of the high resolution images, we did not manage to fit a 4K image into a single GPU during training. Therefore, following recent work on depthwise separable convolutions [18, 7], we replaced all convolutions in a conventional hourglass network [30] with depthwise separable convolutions. This resulted in lowering the number of weights in the network by a factor of 2 and enabled training with 4K images. We refer to the version of the architecture present in Fig. 1 with depthwise separable convolutions as the *light* variant of our network. For resolutions of up to 2048 x 2048, this change wasn't necessary. The effect of introducing depthwise separable convolutions as opposed to standard convolutions into our architecture is analyzed in detail in Section 4.

For all experiments reported in this paper, we use a learning rate of $1e^{-4}$, and lowered it to $1e^{-5}$ after 30 epochs. Models were trained with batch sizes mentioned in Table 1. All models were trained until convergence using the ADAM optimizer [24] on a single NVIDIA 1080Ti GPU. We used pytorch [31] to implement our architecture.

4. Results and Discussion

4.1. Learning Latent Heatmaps

One of the ways in which our approach differs from existing methods in facial landmark detection is the representation of landmarks with learned latent heatmaps. In Fig. 3, we show the differences in the heatmaps produced by the global and different regional models. For the purpose of visualization, the heatmaps predicted by the global model were upscaled using nearest neighbour interpolation and shown alongside the heatmaps predicted by the regional networks. As expected, the global heatmaps are of lower quality but capture the overall structure of the person's face, and therefore result in expected crops. The final column of Fig. 3 shows the precise high resolution heatmaps produced by the regional models. Positions with strong activations in both the global and regional heatmaps indicate the more salient landmarks on the face.

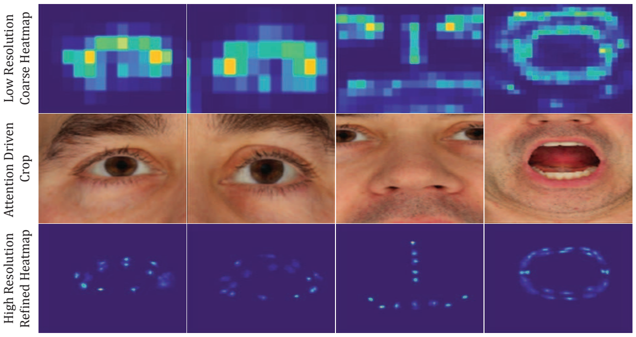
Low Resolution Coarse Heatmap
Attention Driven Crop
High Resolution Refined Heatmap

Figure 3. Manually cropped global heatmaps, corresponding attention-driven crops, and regional high quality heatmaps. Note the precision of the high quality heatmaps, e.g. one can easily distinguish between outer and inner lip landmarks.

4.2. Benefits of Attention-Driven Cropping

Training regional networks on local crops provides several advantages. The first is that it encourages each regional network to concentrate only on a specific region of the face and therefore learn region-specific features that help in predicting landmarks with higher accuracy. Fig. 3 also shows how the coarse global heatmaps of each region are refined by the regional networks. As one would intuitively expect, facial features that are harder to distinguish at lower resolutions start to separate out in the regional heatmaps, resulting in precise localization (Fig. 4). This is especially visible in case of the mouth where the outer and inner lips are clearly separated in the high resolution regional heatmap.

The second advantage is that since each regional model is looking only at specific part of a face, the quality of regional landmarks is independent of the appearance of other regions. We expect that this property of our architecture makes the quality of overall landmark prediction much more robust to global changes in appearance.

Thirdly, our global-local architecture is designed to process only meaningful regions of a high resolution input image. Purposefully discarding irrelevant portions of a high resolution image avoids the need for networks to be extremely deep or build huge feature representations that don't fit inside current GPUs. Concentrating only on RoIs enables the regional hourglasses to leverage the high frequency detail present in the captured imagery by only predicting landmarks within a meaningful ROI. Our approach allows us to perform deep landmark detection on high resolution images, without sacrificing batch size, while at the same time avoiding unnecessary computations.

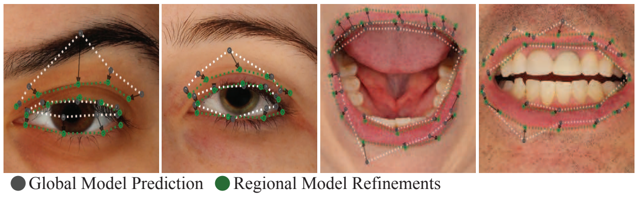

Figure 4. Effect of regional refinement: Local corrections (green) made by regional models to landmarks predicted by the global model (gray) are shown. When provided with sufficient context, regional models can produce both small and large corrections.

Method	Common	Challenging	Full Set
MDM [40]	4.83	10.14	5.88
Two-Stage$_{GT}$ [28]	4.36	7.42	4.96
RDR [43]	5.03	8.95	5.80
FHR [38]	N/A	N/A	3.8
SAN [9]	3.34	**6.6**	3.98
DSRN [29]	4.12	9.68	5.21
TS [10]	2.91	5.91	**3.49**
ODN [54]	3.56	**6.67**	4.17
Ours	**2.83**	7.04	4.23
Ours (51 Hi-res landmarks only)	2.41	5.68	3.50

Table 2. Performance on the 300W dataset. Despite being designed for high resolution imagery, our method performs very well also on low resolution in-the-wild images.

4.3. Evaluations on 300W and 300VW

Though our method was primarily designed for high resolution images, we evaluate our attention driven cropping on the low resolution 300W and 300-VW datasets. For 300W, we split the *Helen*, *LFPW*, *AFW*, and *Ibug* datasets into training and test sets identical to previous methods [9, 29, 38]. We train our model at a resolution of 256x256 pixels and crop sizes of 128x128 pixels, and define 4 regions of interest (2) from which a total of 51 high resolution landmarks are detected and the remaining 17 landmarks on the jawline are predicted by the global hourglass. We report the normalized mean error (NME) [4] metric on the 300-W test sets in Table 2. Even at 256x256 pixels, our method establishes a new baseline on the common subset and remains competitive to state-of-the-art on the challenging subset. Qualitative landmark predictions on the 300W test set are shown in Fig. 5 Additionally, as we will see from the experiments in Section 4.4, the benefits of our attention driven cropping method become significantly larger as we move to higher resolutions.

To validate our method on the 300-VW dataset, we retrain another network identical to the one used for the 300W, using 50 training videos from 300-VW. Table 3 compares our method to existing state-of-the-art using the NME metric on three different test categories. Our method again produces the best results on 2 out of 3 of the categories.

4.4. Evaluations at Higher Resolutions

We compare our attention-driven cropping architecture with a random forest algorithm [23], a two stage hourglass network [30] and a 4 stage hourglass network namely the 2D landmark detector referred to as FAN [4]. We used our low resolution 256 dataset described in Section 3.3 to train the random forest, the 2 and the 4 stage hourglass networks.

Since our architecture enables training with resolutions

Method	Category 1	Category 2	Category 3
SDM [44]	7.41	6.18	13.04
TSCN [35]	12.54	7.25	1.13
CFSS [55]	7.68	6.42	13.67
TCDCN [52]	7.66	6.77	14.98
TSTN [1]	5.36	4.51	12.84
DSRN [29]	5.33	4.92	8.85
FHR+STA [38]	4.40	4.16	**5.96**
Ours	**4.17**	**3.89**	7.28
Ours (51 high res landmarks only)	3.66	3.35	6.65

Table 3. Performance on the 300-VW dataset. Similar to Table 2, our method also performs very well on the videos of 300-VW.

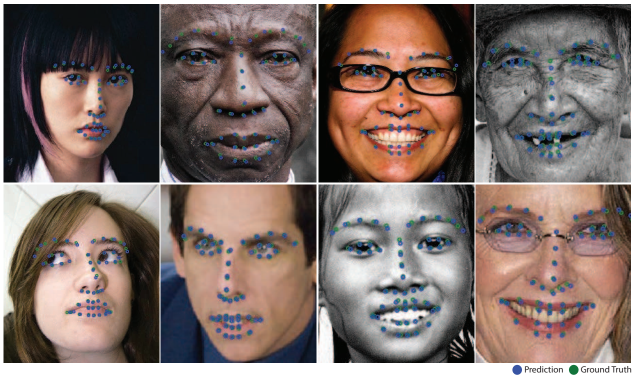

Figure 5. Qualitative results showing attention refined regional landmarks on a few samples from the 300-W test set.

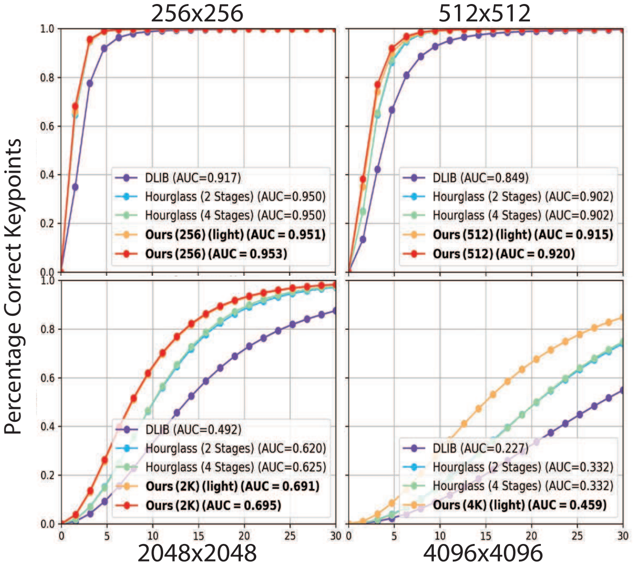

Figure 6. Percentage Correct Keypoints as function of the error in pixels for our method compared to DLIB [23], hourglass [30] and FAN [4] for different resolutions from 256 to 4K. We show that both our regular and 'light' variants outperform previous methods. At 4K, only the 'light' version is possible but still provides significant improvement. Refer to Table 4 for Normalized Mean Errors.

Method	256	512	2048	4096
DLIB [23]	3.72	3.43	3.32	3.35
Hourglass (2 Stages) [30]	2.34	2.34	2.34	2.38
Hourglass (4 Stages) [4]	2.39	2.39	2.39	2.44
Ours (light)	**2.34**	**2.08**	**1.97**	**1.95**
Ours	**2.26**	**1.95**	**1.94**	**-**

Table 4. Normalized Mean Errors for our method compared to DLIB [23], hourglass [30] and FAN [4] for different resolutions from 256 to 4K. Refer to Fig. 6 for Percentage Correct Keypoints visualization.

of up to 4096, we trained it with data of appropriate resolution. We use the high quality test set described in Section 3.3 consisting of 4416 images for evaluation. For the sake of comparison, the predictions made by the random forest, the 2 and the 4 stage hourglasses were up-scaled manually from 256 to the evaluation resolution.

To quantitatively compare landmark predictions, we use the Percentage Correct Keypoints (PCK) metric used by [15] and the Normalized Mean Error (NME) as before. Fig. 6 and Table 4 show quantitative comparisons of our algorithm against different methods at resolutions ranging from 256 to 4096. At a resolution of 4096, we report the PCK and NME metric only for the *light* variant of our architecture for reasons explained in Section 3.4. The resolution of 1024 is considered separately in our ablation study in Section 4.5.

Our approach, including the *light* variant, outperforms other methods across all resolutions, indicating that attention-driven cropping is not only a way for training with higher resolution imagery, but is an effective method for facial landmark detection in principle. The benefits of our approach increase as the resolution of the input increases. This can be inferred from the differences in the area under curve metric between our method and the 4 stage hourglass as the resolution increases.

4.5. Ablation Studies

We evaluate some of our architectural choices at a resolution of 1024. These results are presented in Fig. 7.

Effect of Additional Stages Stacking models on top of one another is a common approach in landmark localization [30, 4, 42]. Such stacking could also be incorporated into our architecture by stacking multiple regional refinement modules on top of one another. When we stack an additional regional hourglass to our base architecture shown in Fig. 1, we see an improvement in the AUC (see Fig. 7, left).

Choice of Architecture The modular nature of our architecture makes it possible to swap the hourglass with different fully convolutional architectures. To validate the robustness of the proposed concept to different architectural choices, we consider two recently proposed fully convolutional architectures i) a 6 stage Convolutional Pose Machine (CPM) [42, 6] and ii) the CNN 6/7 architecture from [13]. When using CNN 6/7, we discard the last fully connected layer to keep the network fully convolutional and append additional CNN 6/7 stages where each stage receives both the image and the heatmap of the previous stage as

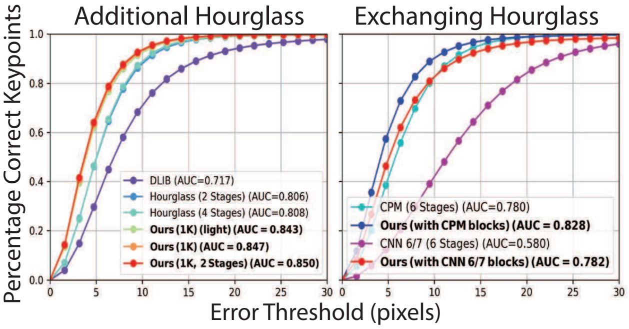

Figure 7. Ablation study at 1K. Left: stacking an additional regional hourglass improves the AUC. Right: swapping the hourglass with a CPM [42, 6] or CNN 6/7 architecture [13] shows that our attention-driven cropping scheme can improve other architectures too, but the hourglass still obtains the best results.

input. We retrain both the 6 stage CPM and the 6 stage CNN 6/7 architecture and compare them with our attention-driven cropping concept where every hourglass module is replaced by a single stage CPM or CNN 6/7 respectively. In the right half of Fig. 7, we see the results of this experiment. Though our attention-driven CPM and attention-driven CNN 6/7 consist of lesser parameters than the 6 stage CPM and 6 stage CNN 6/7 architectures respectively, we see a large improvement in switching to landmark detection with our attention-driven cropping concept as opposed to holistic multi-stage methods. This superior performance is a testament to the robustness of the proposed method and its applicability to more general problems in localization.

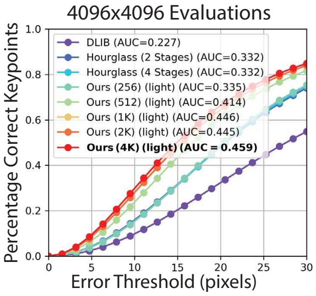

Figure 8. Comparing results of testing at 4K but training at different resolutions confirms there is indeed a benefit by moving to higher resolution training when possible.

4.5.1 Necessity of High Resolution Detection

The 4K images we captured (Section 3.3) were annotated by a human expert. Considering the limited precision with which humans annotate landmarks [11], there is a question over the necessity of training a model with extremely high resolution imagery. In Fig. 8, we compare the results of our *light* variant with hourglass building blocks, trained at 4096 to up-scaled predictions from other attention driven models trained at lower resolution. The performance of the attention-driven cropping framework increases as the resolution of the input data increases, ultimately making the model directly trained at 4K resolution the best performing model. From this, we see that there is indeed a benefit by moving to higher resolutions whenever possible. In Fig. 9, we show qualitative results on a few different test images.

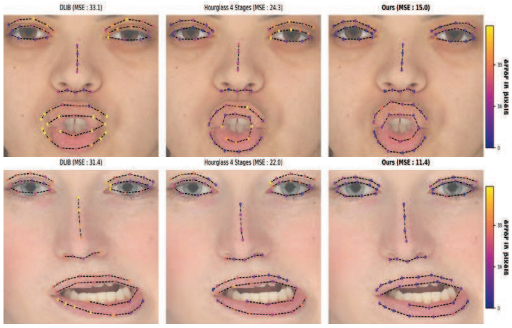

Figure 9. Qualitatively, our method produces the most accurate landmarks on a test image set. Here we compare to DLIB [23] and a 4-stage hourglass (FAN) [4] on a small set of the test data. Pixel errors are indicated by color.

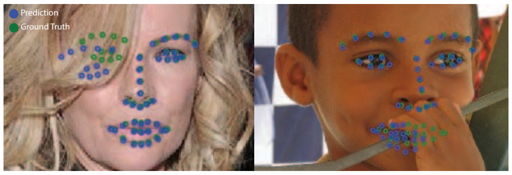

Figure 10. Situations where our regional refinement could fail are shown here on the 300w dataset, where the crop results in meaningless images when parts of the face are completely occluded. In such cases, a global approach would be more preferable.

4.6. Limitations

The proposed method is designed to improve landmark localization by leveraging information present at higher resolutions. If no additional information is present, or the additional information is deceiving as is the case for partial occlusions (Fig. 10), the performance degrades. This is related to the classical aperture problem, and future work could investigate approaches to determine the best resolution to localize features in automatically.

5. Conclusion

We present a novel, fully convolutional regional architecture designed to predict landmarks on very high resolution images. Our proposal is an end-to-end attention-driven architecture that allows to train deep networks on higher resolution images by automatically defining and focusing on regions of interest instead of considering the image holistically. We show that our architecture achieves superior performance over holistic state of the art convolutional architectures across all resolutions from 256 to 4K. We believe our method fills the need for algorithms that can leverage the rich information available in high resolution imagery, which is becoming increasingly more common.

References

[1] Two-stream transformer networks for video-based face alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:2546–2554, 2018. 2, 7

[2] Thabo Beeler, Bernd Bickel, Paul A. Beardsley, Bob Sumner, and Markus H. Gross. High-quality single-shot capture of facial geometry. *ACM Trans. Graph.*, 29(4):40:1–40:9, 2010. 5

[3] Chandrasekhar Bhagavatula, Chenchen Zhu, Khoa Luu, and Marios Savvides. Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. *CoRR*, abs/1707.05653, 2017. 2

[4] 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). *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1021–1030, 2017. 2, 3, 6, 7, 8

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

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1611.08050, 2016. 3, 7, 8

[7] François Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016. 5

[8] Jiankang Deng, George Trigeorgis, Yuxiang Zhou, and Stefanos Zafeiriou. Joint multi-view face alignment in the wild. *CoRR*, abs/1708.06023, 2017. 2

[9] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. *CoRR*, abs/1803.04108, 2018. 2, 6

[10] Xuanyi Dong and Yezhou Yang. Teacher supervises students how to learn from partially labeled images for facial landmark detection. *ArXiv*, abs/1908.02116, 2019. 2, 6

[11] 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. *CoRR*, abs/1807.00966, 2018. 8

[12] Yuan Dong and Yue Wu. Adaptive cascade deep convolutional neural networks for face alignment. *Computer Standards and Interfaces*, 42:105–112, 2015. 2

[13] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiaojun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. *CoRR*, abs/1711.06753, 2017. 2, 5, 7, 8

[14] Ross B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015. 4

[15] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587, 2014. 4, 7

[16] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 9:249–256, 13–15 May 2010. 5

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988, 2017. 2, 4

[18] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. 5

[19] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5d heatmap regression. *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, pages 125–143, 2018. 2, 3, 4

[20] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via cnn-based dense 3d model fitting. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4188–4196, 2016. 2

[21] Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. Pose-invariant face alignment with a single CNN. *CoRR*, abs/1707.06286, 2017. 2

[22] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. 5

[23] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014. 2, 6, 7, 8

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 5

[25] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. *CoRR*, abs/1706.01789, 2017. 2

[26] Zhujin Liang, Shengyong Ding, and Liang Lin. Unconstrained facial landmark localization with backbone-branches fully-convolutional networks. *CoRR*, abs/1507.03409, 2015. 2

[27] Yaojie Liu, Amin Jourabloo, William Ren, and Xiaoming Liu. Dense face alignment. *CoRR*, abs/1709.01442, 2017. 2

[28] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 2, 6

[29] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. Direct shape regression networks for end-to-end face alignment. In *CVPR*, 2018. 2, 6, 7

[30] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 483–499, 2016. 2, 3, 5, 6, 7

[31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[32] Rajeev Ranjan, Vishal M. Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *CoRR*, abs/1603.01249, 2016. 2

[33] 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. *2015 IEEE International Conference on Computer Vision Workshop, ICCV Workshops 2015, Santiago, Chile, December 7-13, 2015*, pages 1003–1011, 2015. 2

[34] 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 *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015. 2

[35] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, pages 568–576, Cambridge, MA, USA, 2014. MIT Press. 2, 7

[36] Keqiang Sun, Wayne Wu, Tinghao Liu, Shuo Yang, Quan Wang, Qiang Zhou, , Zuochang Ye, and Chen Qian. Fab: A robust facial landmark detection framework for motion-blurred videos. In *ICCV*, 2019. 2

[37] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 3476–3483, Washington, DC, USA, 2013. IEEE Computer Society. 2

[38] Ying Tai, Yicong Liang, Xiaoming Liu, Lei Duan, Ji-Lin Li, Chengjie Wang, Feiyue Huang, and Yu Chen. Towards highly accurate and stable face alignment for high-resolution videos. *CoRR*, abs/1811.00342, 2018. 2, 6, 7

[39] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. *CoRR*, abs/1411.4280, 2014. 2

[40] George Trigeorgis, Patrick Snape, Mihalis A. Nicolaou, Epameinondas Antonakos, and Stefanos P. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4177–4187, 2016. 2, 6

[41] Georgios Tzimiropoulos. Project-out cascaded regression with an application to face alignment. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3659–3667, 2015. 2

[42] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4724–4732, 2016. 2, 3, 7, 8

[43] Shengtao Xiao, Jiashi Feng, Luoqi Liu, Xuecheng Nie, Wei Wang, Shuicheng Yan, and Ashraf A. Kassim. Recurrent 3d-2d dual learning for large-pose facial landmark detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1642–1651, 2017. 6

[44] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 532–539, Washington, DC, USA, 2013. IEEE Computer Society. 2, 7

[45] Xiang Jun Xu and Ioannis A. Kakadiaris. Joint head pose estimation and face alignment framework using global and local cnn features. *2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)*, pages 642–649, 2017. 2

[46] J. Yang, Q. Liu, and K. Zhang. Stacked hourglass network for robust facial landmark localisation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2025–2033, July 2017. 2

[47] Xiang Yu, Feng Zhou, and Manmohan Chandraker. Deep deformation network for object landmark localization. *CoRR*, abs/1605.01014, 2016. 2

[48] Stefanos P. Zafeiriou, George Trigeorgis, Grigorios Chrysos, Jiankang Deng, and Jie Shen. The menpo facial landmark localisation challenge: A step towards the solution. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2116–2125, 2017. 2

[49] Jie Zhang and Robert B. Fisher. 3d visual passcode: Speech-driven 3d facial dynamics for behaviometrics. *Signal Processing*, 160, 02 2019. 2

[50] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *CoRR*, abs/1604.02878, 2016. 2

[51] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *In ECCV. 94–108*, 2014. 2

[52] 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*, 38:918–930, 2016. 2, 7

[53] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops*, ICCVW '13, page 386–391, USA, 2013. IEEE Computer Society. 2

[54] Meilu Zhu, Daming Shi, Mingjie Zheng, and Muhammad Sadiq. Robust facial landmark detection via occlusion-adaptive deep networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 6

[55] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4998–5006, 2015. 2, 7

[56] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3d solution. *CoRR*, abs/1511.07212, 2015. 2