Fine-grained facial landmark detection exploiting intermediate feature representations

Yongzhe Yan, Stefan Duffner, Priyanka Phutane, Anthony Berthelier, Xavier Naturel, Christophe Blanc, Christophe Garcia, Thierry Chateau

Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Fine-grained facial landmark detection exploiting intermediate feature representations

Yongzhe Yan[a,c,**], Stefan Duffner[b], Priyanka Phutane[c], Anthony Berthelier[a,c], Xavier Naturel[c], Christophe Blanc[a], Christophe Garcia[b], Thierry Chateau[a]

[a] *Université Clermont Auvergne, CNRS, SIGMA, Institut Pascal, F-63000 Clermont-Ferrand, France*
[b] *Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France*
[c] *Wisimage, France*

ABSTRACT

Facial landmark detection has been an active research subject over the last decade. In this paper, we present a new approach for Fine-grained Facial Landmark Detection (FFLD) improving on the precision of the detected points. A high spatial precision of facial landmarks is crucial for many applications related to aesthetic rendering, such as face modeling, face animation, virtual make-up, etc. In this paper, we present an approach that improves the detection precision. Since most facial landmarks are positioned on visible boundary lines, we train a model that encourages the detected landmarks to stay on these boundaries. Our proposed Convolutional Neural Networks (CNN) effectively exploits lower-level feature maps containing abundant boundary information. To this end, beside the main CNN predicting facial landmark positions, we use several additional components, called CropNets. CropNet receives patches cropped from feature maps at different stages of this CNN, and estimate fine corrections of its predicted positions. We also introduce a novel robust spatial loss function based on pixel-wise differences between patches cropped from predicted and ground-truth positions. To further improve the landmark localisation, our framework uses several loss functions optimising the precision at several stages in different ways. Extensive experiments show that our framework significantly increases the local precision of state-of-the-art deep coordinate regression models.

1. Introduction

Facial landmark detection algorithms aim to retrieve the coordinates of a number of characteristic points given a face image. It has become an important step in various tasks such as facial expression analysis (Martinez et al., 2017), 3D face reconstruction (Jackson et al., 2017) and face recognition (Ding and Tao, 2016). In this work, we are interested in *Fine-grained Facial Landmark Detection* (FFLD), which is required for applications such as face modeling, virtual make-up, and in tasks that require pixel-level accuracy for aesthetic rendering. In these applications, slight displacements of the estimated landmark positions significantly deteriorate the user experience. In existing methods from the literature, this refinement is usually performed in a post-processing step (Zeng et al., 2015; Wang et al., 2017).

Recently, deep Convolutional Neural Networks (CNN) have brought a remarkable improvement in this regard. We can categorize these deep CNN models into two different types according to their network output: *Coordinate Regression Models* (CRMs) and *Heat-map Regression Models* (HRMs) (Wu and Ji, 2017; Yan et al., 2018; Nibali et al., 2018; Merget et al., 2018). Most of the CRMs end with a Fully-Connected (FC) layer to directly predict the numeric coordinate values. On the other hand, HRMs generally adopt a Fully Convolutional Neural Network that provides one "heat map" per landmark as output. Each pixel value of a particular heat map represents the conditional probability of the landmark being present at this position. HRMs are an alternative to CRMs, and have become popular in recent years due to their strong capacity of handling large head poses. However, both of the approaches cannot directly provide highly accurate FFLD due to the following reasons.

Local imprecision problem of CRMs: A single-stage CRM usually suffers from local imprecision. This is likely due to the loss of local detail through successive feature map down-

**Corresponding author:
e-mail: yongzhe.yan@etu.uca.fr (Yongzhe Yan)

sampling. Hence, numerous coarse-to-fine methods (Sun et al., 2013; Zhou et al., 2013; Zhang et al., 2014a; Trigeorgis et al., 2016; Fan and Zhou, 2016; Kowalski et al., 2017; Chen et al., 2017; He et al., 2017b; Lv et al., 2017) have been proposed to cope with this issue. In most of them, the refinement is performed in a cascade that sequentially processes local image patches to recover the local detail information.

Local imprecision problem of HRMs: The imprecision of HRMs is mainly due to quantization error in the output heat maps. In most of these models, the final landmark prediction is obtained from the position of the maximum value in the output heat map, thus an integer value. Furthermore, the predicted heat map is typically around four times smaller than the input image. These two factors cause considerable quantization errors for HRMs.

To enable FFLD, we propose to combine the advantages of both CRMs and HRMs. We propose to exploit the intermediate low-level feature maps in CRMs and reuse them in an additional processing step. This reuse of low-level feature maps is inspired by skip connections that are widely used in HRMs (with encoder-decoder structures). It enables local detail information to be directly used by higher-level processing stages in the neural network. We found that the boundaries of facial components still remain clear on the low-level feature maps, even if the CRM output suffers from the local imprecision problem (see Fig. 2). This shows the potential that our method leverages by using this information in the output layers. On the other hand, since we adopt a CRM framework, the output predicted by the final FC layer is a vector of real (floating-point) numbers, which avoids the quantization errors inherent in HRMs.

Besides, the $L2$ loss used to train HRMs (Newell et al., 2016; Bulat and Tzimiropoulos, 2017b) differs from the traditional one commonly used for CRMs. The heat map $L2$ loss computes the pixel-wise $L2$ distance between the predicted heat maps and target heat maps. It is robust to outliers, as its value saturates when two Gaussian distribution (representing the ground-truth and predicted landmark positions) do not overlap regardless of the distance between two landmarks. More robust loss functions have proved to be helpful for CNNs to focus on small-range errors (Belagiannis et al., 2015; Feng et al., 2018), and we will show that this is beneficial for FFLD.

Generally speaking, our CRM-based coarse-to-fine framework (see Fig. 1) is inspired by the use of skip connections and the robust spatial loss function used in HRMs. The main contributions of this paper are:

- A novel *feature map patch alignment* method (Sect. 3), to establish skip connections in coordinate regression models, where small subsidiary networks, called *CropNet*s, cooperates with the main CNN (baseline network) to provide small corrections. These CropNets leverage local detail information from local patches in low-level feature maps. The refined correction is based on a direct measure of the crop misalignment. Unlike the previous coarse-to-fine methods, our baseline network can be jointly learned with refinement since CropNets enable the gradient to be directly back-propagated through the low-level feature maps.
- A novel robust loss function named *Align Loss* (Sect. 4.3) that measures the minor but important misalignment between the patches cropped by the ground truth and predicted localization. This loss calculates the pixel-wise value differences between two patches. Compared to standard $L2$ Loss, our Align Loss forces the predicted landmarks to stay on boundary lines and is thus able to improve the precision to the pixel level.
- A *multi-loss training scheme* (Sect. 4), where different loss functions in different refinement stages are employed. To achieve extreme localization precision, loss functions sensitive to big errors are assigned to coarser stages, and loss functions more sensitive to small errors are assigned to finer stages.

2. Related Work

CRMs: Using deep CRMs to predict facial landmark location has gained popularity after the success of the ImageNet challenge (Krizhevsky et al., 2012). Some of the methods (Zhang et al., 2014b; Ranjan et al., 2017; Honari et al., 2018) refine the facial landmarks by using auxiliary attributes while most of the methods adopt a coarse-to-fine detection. Sun et al. (2013); Zhang et al. (2014a); Zhang and Hu (2018) used cascaded CNNs to refine the landmark regression. Chen et al. (2017) proposed component-wise and point-wise refinement stages to rectify the landmark locations step-by-step. Trigeorgis et al. (2016) proposed to use an RNN as regressor to perform a cascaded-regression-like coarse-to-fine landmark detection. Recently, the Spatial Transformer Network (Jaderberg et al., 2015) has been adopted to help coarse-to-fine deep facial landmark regression (Yu et al., 2016; Lv et al., 2017). Dong et al. (2018b) proposed an unsupervised approach to refine the video landmark detection by optical flow coherency.

HRMs: HRMs establish an image-to-image mapping between input images and heat maps where each landmark is represented as a 2D (discretized) Gaussian distribution. They have been originally introduced by (Duffner and Garcia, 2005) and gained much popularity recently. Usually, a symmetric structure is adopted and sometimes skip connections are used which enable low-level information to flow from the encoder directly to the decoder. This model has been widely-used for human pose detection (Wei et al., 2016; Newell et al., 2016) and later proved to be efficient for facial landmark detection as well (Bulat and Tzimiropoulos, 2017a,b; Chen et al., 2019; Lai et al., 2018; Dong et al., 2018a; Valle et al., 2018). One disadvantage of HRMs is that the final prediction is obtained by taking the location with the maximum pixel value. This operation brings inevitable quantization errors. To mitigate this issue, Sun et al. (2018) proposed an integral regression method where the landmark coordinates are estimated as the integration of all probabilities over the heat map. Tai et al. (2019) introduced a fractional heat map regression approach to tackle this issue. Wu et al. (2018) proposed to predict heat maps of facial component boundaries and used this boundary information to improve the landmark coordinate regression. Finally, Tang et al. (2018)

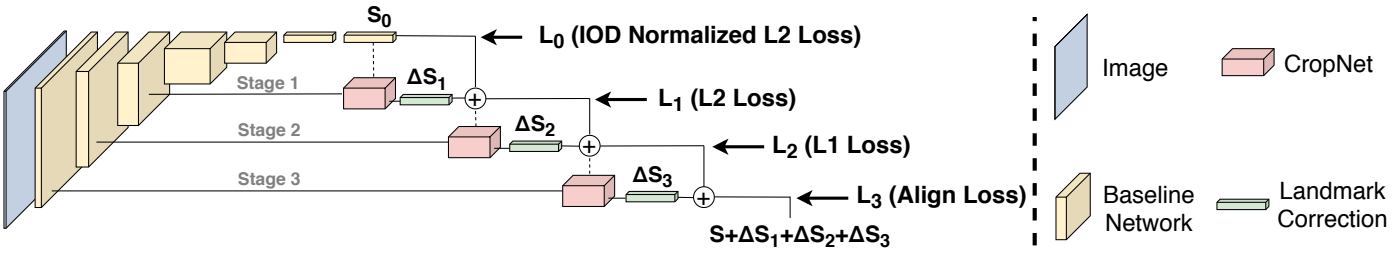


Fig. 1: Overview of our 3-stage coarse-to-fine coordinate regression framework. It contains skip connections between the intermediate feature maps and the main network output S_0. In each stage, a CropNet (described in Sect. 3.2) refines the landmark location (ΔS) based on the patches cropped from lower-level feature maps. The refined landmark locations are passed to the next stage (dotted lines) for cropping. Different loss functions (IOD Normalized $L2$ Loss, $L2$ Loss, $L1$ Loss and Align Loss described in Sect. 4.3) are used to train different refinement stages in an end-to-end manner (cf. details of gradient back-propagation in Sect. 5).

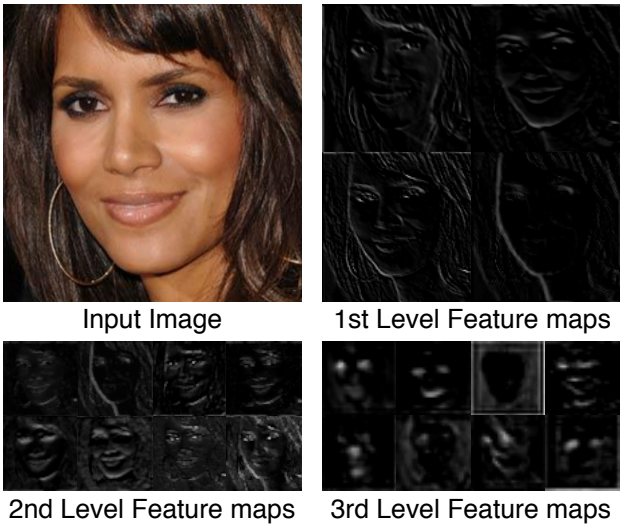


Fig. 2: A visualization of several feature maps in different levels of ResNet18 trained for facial landmark detection. We can observe that low level feature maps retain abundant visual boundary information that we exploit in our approach. Higher-level feature maps present more general information compared to lower-level ones. (The spatial dimension of the third-level feature maps is increased and interpolated for better visibility.)

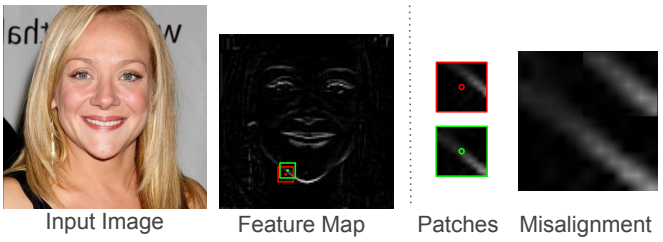


Fig. 3: The main idea of feature map patch alignment: A misplaced landmark (in red) leads to a boundary misalignment on the cropped feature map patch compared to the ground truth (in green). Our model measures this patch misalignment to estimate a refined correction to coarse landmark prediction. Note that our patch alignment approach is different from existing image alignment methods which take pairs of input images (Chang et al., 2017): it only uses misaligned patches as input and learns misalignment for each landmark based on the statistics.

presented a method with quantized densely connected U-Nets, which greatly improved the efficiency of HRMs.

Deep Robust Regression: Robust training is critical to enhance the landmark accuracy especially for small errors. Previous work on robust loss function for deep model regression is mainly inspired by the use of the M-estimator in robust statistics. The primary goal is to attenuate the impact of outliers on the overall loss. Belagiannis et al. (2015) proposed to use Tukey's biweight loss function for human pose estimation. Their loss function saturates with large residuals. They showed that their loss function helps the deep regression model to converge both faster and better, compared to the traditional $L2$ loss function. Feng et al. (2018) proposed a novel wing loss for deep robust facial landmark detection, which behaves like the logarithmic function for small errors and like the $L1$ loss function for large errors. They emphasized the importance of small residuals during the calculation of the loss. Recently, Lathuilière et al. (2018) combined a robust mixture modeling to deep CNN regression models which adapts to an evolving outlier distribution without setting a manual threshold.

3. Feature Map Patch Alignment

In HRMs, the skip connections are intuitive due to the similar spatial dimension shared between input layers and output layers in each stage. However, in CRMs, the spatial resolutions do not match. The dimension of low-level feature maps are too large for the output FC layer. Miao et al. (2018) and Yue et al. (2018) used a non-linear embedding layer to reduce the dimension of feature maps for the succeeding FC layer. In contrast, we propose to reduce the feature map dimension by cropping patches around each landmark. Therefore, we developed a feature map patch alignment method (see Fig. 3) for FFLD described in the following sections.

3.1. Baseline Network

The aim of facial landmark coordinate regression is to establish a non-linear mapping between an image $\mathcal{I} \in \mathbb{R}^{h \times w}$ and landmark Cartesian coordinates $\mathcal{S} \in \mathbb{R}^{2N}$, where N represents the number of landmarks. We use ResNet (He et al., 2016) as our baseline network but reduce 75% of its feature map channels in each layer. Despite a very good overall prediction performance of this model, it lacks some local precision. Following refinement is then performed based on this baseline network.

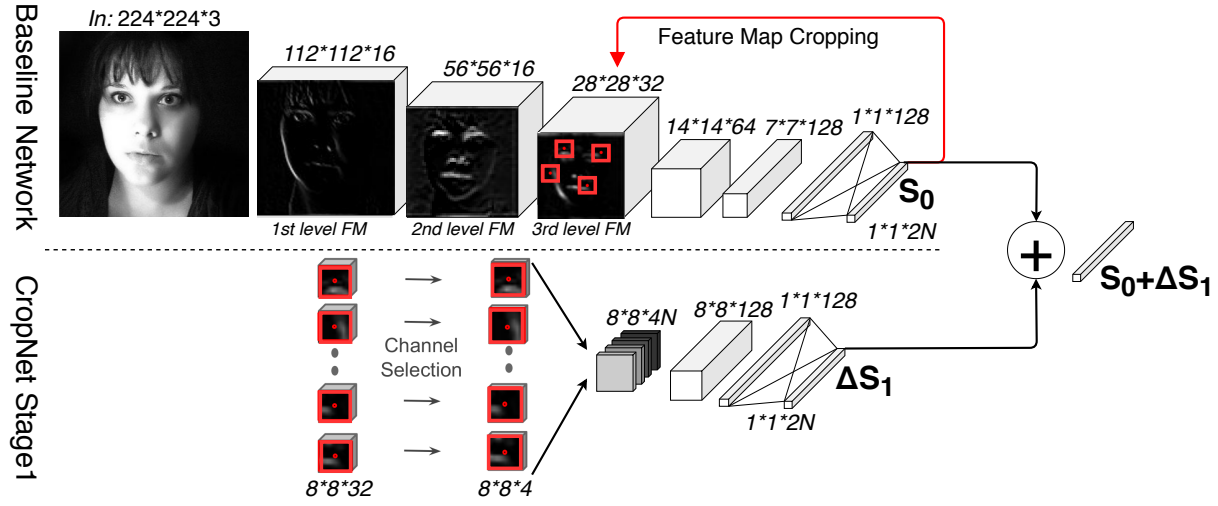


Fig. 4: An illustration of our feature map patch alignment method in the first stage. Small patches are cropped from the 3rd-level feature maps based on the coarse landmark detection S_0 given by the baseline network. The number of channels is reduced by a linear 1×1 convolutional layer. The selected feature maps are then concatenated as input to the CropNet, which predicts the correction of landmark localization ΔS_1. Finally, $S_0 + \Delta S_1$ is passed as coarse prediction to crop patches from the 2nd-level feature maps in the next stage.

3.2. CropNet

The objective of the CropNet modules is to find a correction to the coarse prediction, based on the low-level feature maps from the main network. Figure 4 illustrates how we utilize CropNet to refine the facial landmark localization in the first refinement stage. In order to efficiently process detailed information on high-dimensional low-level feature maps, the input dimension of the refinement network is reduced. Given low-level feature maps of dimension (C, H, W) as input, we propose to reduce the dimension of (H, W) by cropping feature maps and to reduce the number of channels C by learning a linear channel selection explained in the following.

Feature map cropping: Similar to previous coarse-to-fine frameworks (He et al., 2017b), we perform a central crop according to the coarse prediction from the previous stage. As shown in Fig. 4, the spatial dimension is reduced from 28×28 to 8×8 in stage 1. We crop patches of 8×8 from all of the 3rd, 2nd and 1st level feature maps in stage 1, stage 2 and stage 3 respectively. In different stages, due to identical patch sizes but different feature map sizes, the patches have different support on the input image. Thus, patches cropped from the 3rd level feature maps have bigger support but contain coarser information while the patches cropped from the 1st level feature maps have smaller support and contain more details. This corresponds well to a coarse-to-fine strategy, where relatively larger errors are corrected in stage 1 and detailed, pixel-level errors in stage 3.

To avoid introducing additional quantization error, the pixels on cropped patches are resampled by bilinear interpolation from original feature maps (as in the Mask-RCNN (He et al., 2017a), for example). This enables us to crop a feature map even on a non-integer position.

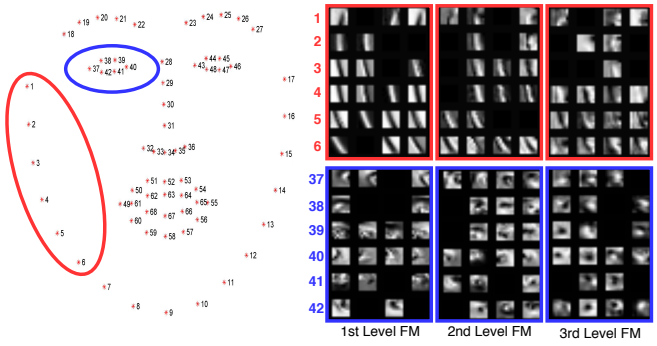


Fig. 5: Examples of the patches around landmarks on the face contour (red, top row) and the eye contour (blue, bottom row) after channel selection. Four channels (columns) are selected per landmark (lines).

Channel selection: Without the reduction of channels, the input channel dimension to CropNet would lead to high computational complexity. Therefore, we propose a learning-based channel selection by using a linear 1×1 convolutional layer. It learns to select and combine the most useful channels (here: 4) from the original ones, especially the ones containing boundary information. As shown in Fig. 4, the channel dimension is reduced from $32N$ to $4N$ in stage 1. Several output examples from channel selection are visualized in Fig. 5. This illustrates that most of the selected channels contain important boundary information.

4. Multi-loss Training Scheme

We propose to use a set of different loss functions to train different stages. Unlike other methods (Lv et al., 2017) that use multiple loss functions for multiple tasks, we use multiple loss functions in different stages to optimize the same overall target. Our motivation is that some loss functions are more adapted to optimize the bigger errors for coarse detection, yet other loss functions are more sensitive to the smaller errors for fine detection.

4.1. IOD Normalized L2 Loss

To train our baseline network, we the $L2$ loss is used, i.e. the squared error of landmark positions, normalized by the Inter-Occular Distance, like (Lv et al., 2017; Kowalski et al., 2017):

$$L = \frac{\|S_{GT} - S_{Pred}\|_2^2}{d}, \tag{1}$$

where S_{GT} and S_{Pred} denote the ground-truth positions (shape) and predicted positions respectively. d denotes the Inter-Ocular Distance. The $L2$ penalizes big errors, especially those occurring in hard examples with large head pose.

4.2. L2 & L1 Loss

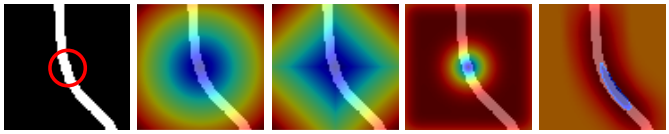
For illustration, in Fig. 6 we visualize the values of different loss functions on a synthetic example. The traditional $L2$ loss (Fig. 6 (b)) used in CRMs calculates the Euclidean distance between the Cartesian coordinates of prediction and ground truth. The loss values grow infinitely when the predictions get further away from the ground truth. For the $L2$ loss function computed pixel-wise on a heat map (Fig. 6 (d)), the loss values saturate when they are far away from the ground truth. Hence, compared to the heat map $L2$ loss function, the standard $L2$ loss is more suitable for minimizing relatively big errors since the large errors do not saturate and thus do not vanish in the gradient descent optimisation. Therefore, we use standard $L2$ loss in the first refinement stage.

Feng et al. (2018) showed that the $L1$ loss function (Fig. 6 (c)) performs better than $L2$ as it focuses on middle and small-ranged errors. Thus, we propose to use the $L1$ loss function in the second refinement stage.

4.3. Align Loss

To provide pixel-level precision and force the predicted landmarks to stay on the boundary lines, we propose a novel loss function, called "Align Loss". The Align Loss is used to train our CropNet in the last refinement stage, i.e. the one with the most detailed feature maps. The Align Loss operation is intuitive when applied on small patches (visualized in Fig. 5). It measures patch misalignment by calculating the pixel value difference between the patches cropped by the ground truth and the prediction.

Our Align Loss is inspired by the robust spatial $L2$ loss used in HRM-based approaches. It calculates the pixel-wise squared difference of values between the patches cropped using the ground truth location and the patches cropped using the predicted location. To facilitate the network convergence, we apply a Gaussian kernel G to filter both of the patches prior to the loss calculation. This essentially smooths the gradient spatially over the patch region, and helps the gradient descent algorithm to converge to the optimal solution.



(a) Patch (b) $L2$ (c) $L1$ (d) HM $L2$ (e) Align

Fig. 6: A synthetic example to illustrate different loss functions for facial landmark detection. We simulate an artificial feature map patch as a crop on a landmark localized on the face contour. (a) is the feature map patch center cropped by the ground-truth landmark location. The red circle indicates the ground-truth landmark location. Each pixel value on (b), (c), (d), (e) represents the loss value when the prediction is positioned on this pixel. Blue indicates lower loss values while red indicates higher values. (b), (c), (d), (e) represent respectively $L2$ loss, $L1$ loss, heat-map regression $L2$ Loss and our Align Loss. Note that the loss values are normalized on each image.

Our loss function L for a given stage can be represented as:

$$L = \sum_{m=1}^{|P_{GT}|} \|(G * P_{GT})_m - (G * P_{S+\Delta S})_m\|_2^2, \tag{2}$$

where P_{GT} indicates the patches cropped by ground-truth locations (with m being the pixel index), and $P_{S+\Delta S}$ indicates the patches cropped by the CropNet prediction. The $*$ refers to a convolution operation.

Our Align Loss bears three advantages:

- It forces the landmark refined by our CropNet to stay on visual boundary lines. It improves landmark precision to the pixel level because even a small patch misalignment may result in large loss values, which is beneficial for FFLD.
- It is more sensitive to the misalignment in the orthogonal direction to the boundary than those along the boundary direction (see 6 (e)). We believe that if a landmark is misaligned along the boundary (e.g. on the chin or cheek contour), it is visually more acceptable than a landmark misaligned in the orthogonal direction even though the error remains the same. This is supported by the work of Dong et al. (2018b), where it has been observed that there exists a random error of manual annotation along the boundary direction.
- It is robust to outliers as they have less influence during the training stage, even if they are out of the patch scope.

To summarize, based on the different characteristics of the loss functions introduced before, we assign them as follows: IOD -normalized $L2$ loss for the baseline network (big errors on hard examples); standard $L2$ for the first refinement stage (relatively big errors); $L1$ loss for the second refinement stage (middle and small-range errors) and Align Loss for the last refinement stage (pixel-level errors).

5. Gradient Backpropagation

Our framework enables the gradient to be back-propagated through both *low-level feature maps* and *crop locations*. This

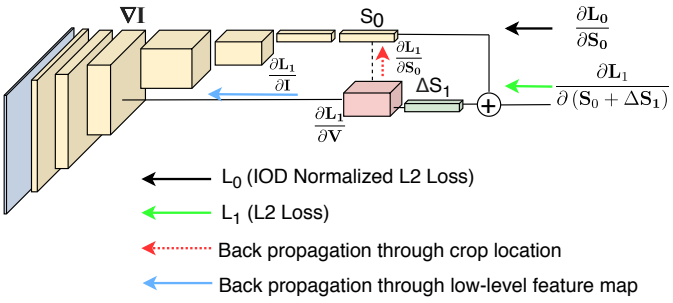


Fig. 7: An illustration of gradient back-propagation in the first refinement stage of our framework.

process is demonstrated in Fig. 7. Deriving the crop operation gives the gradient with respect to: (1) low-level feature maps (blue arrow) and (2) crop location (red dotted arrow).

In this section we will show the composition of the overall gradient ∇I on the low-level feature map (shown in the upper left of Fig. 7). Following Jaderberg et al. (2015) and He et al. (2017b), we demonstrate an example on the first refinement stage, but it is similar in all other refinement stages.

5.1. Preliminaries

First, we compute the gradient of loss L_1 (green arrow in Fig. 7) w.r.t the cropped feature map patch V (red cube in Fig. 7):

$$\begin{aligned}\frac{\partial L_1}{\partial V} &= \frac{\partial L_1}{\partial (S_0 + \Delta S_1)} \cdot \frac{\partial (S_0 + \Delta S_1)}{\partial \Delta S_1} \cdot \frac{\partial \Delta S_1}{\partial V} \\ &= \frac{\partial L_1}{\partial (S_0 + \Delta S_1)} \cdot \frac{\partial \Delta S_1}{\partial V},\end{aligned} \tag{3}$$

where $\frac{\partial L_1}{\partial (S_0+\Delta S_1)}$ is the gradient of the loss L_1 w.r.t. the output coordinates of the first refinement stage, and $\frac{\partial \Delta S_1}{\partial V}$ is the gradient of the CropNet output w.r.t. its input (standard CNN back-propagation).

Given $\frac{\partial L_1}{\partial V}$, we derive the gradient through our crop operation Γ, which can be defined as:

$$V = \Gamma(I, S_0), \tag{4}$$

where I is the low-level feature map, and S_0 is the crop location obtained by the baseline network. We will derive the gradient w.r.t. both I : $\frac{\partial L_1}{\partial I}$ (blue arrow in Fig. 7) and S_0 : $\frac{\partial L_1}{\partial S_0}$ (red dotted arrow in Fig. 7) respectively in Sect. 5.2 and Sect. 5.3.

Considering that bilinear interpolation is used when we crop the patches, the pixel positioned at (q, p) of V is obtained as:

$$V_{qp} = \sum_{n=0}^{H-1} \sum_{m=0}^{W-1} I_{nm} \max\left(0, 1 - \left|y_q - n\right|\right) \max\left(0, 1 - \left|x_p - m\right|\right) \tag{5}$$

$$y_q = y - (h-1)/2 + q \tag{6}$$

$$x_p = x - (w-1)/2 + p, \tag{7}$$

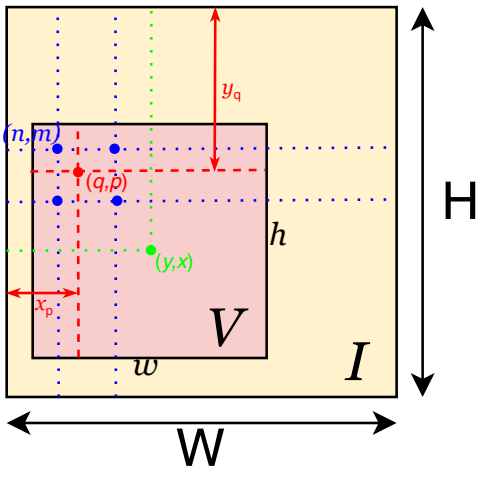


Fig. 8: An illustration of the crop operation with bilinear interpolation.

where $(y_q, x_p) \in \mathbb{R}^2$ is the corresponding position of (q, p) w.r.t the entire feature map I, and (y, x) (components of S_0) is the crop location of each landmark. h and w represent the height and the width of the patch V. H and W represent the height and the width of the feature map I. I_{nm} is the value of the pixel positioned at (n, m) on I.

The previous process is more intuitively illustrated in Fig. 8. The feature map patch V (red square) is center-cropped on the low-level feature map I (yellow square) at position (y, x) (green point). Note that y and x are floating-point numbers. Therefore, the pixels on the patch V (e.g. the red point) are not necessarily aligned to the pixels on the low-level feature maps I (e.g. the blue points). With bilinear interpolation, the value of the pixel V_{qp} on the feature map patch (red point) is determined by the values of its 4 neighbouring pixels on the low-level feature maps (blue points), based on their values and distance to the red point.

The term $\max\left(0, 1 - \left|y_q - n\right|\right) \max\left(0, 1 - \left|x_p - m\right|\right)$ in Eq. 5 ensures that only the four neighbouring pixels influence V_{qp}, based on their distance to (q, p). Eq. 5 was intentionally written in this way for easier presentation in Sect. 5.3.

5.2. Gradient w.r.t. low-level feature maps

We derive the gradient w.r.t the low-level feature maps.

$$\nabla_{low\_level} = \frac{\partial L_1}{\partial I} = \frac{\partial L_1}{\partial V} \cdot \frac{\partial V}{\partial I}. \tag{8}$$

Thus, for each (n, m) on I and using Eq. 5, we have:

$$\begin{aligned}\frac{\partial L_1}{\partial I_{nm}} &= \sum_{p,q} \frac{\partial L_1}{\partial V_{qp}} \cdot \frac{\partial V_{qp}}{\partial I_{nm}} \\ &= \sum_{p,q} \frac{\partial L_1}{\partial V_{qp}} \max\left(0, 1 - \left|y_q - n\right|\right) \max\left(0, 1 - \left|x_p - m\right|\right).\end{aligned} \tag{9}$$

The value of y_q and x_p can be obtained from S_0, h and w in Eq. 6 and Eq. 7. We can therefore calculate $\frac{\partial L_1}{\partial I}$ in Eq. 8

since $\frac{\partial L_1}{\partial V_{qp}}$ has been obtained in Eq. 3. In fact, this step can be intuitively interpreted as a reprojection of the gradient on V back to the corresponding position on I based on the crop location.

5.3. Gradient w.r.t. crop locations

We now derive the gradient of the loss function w.r.t. a crop location $\frac{\partial L_1}{\partial S_0}$:

$$\frac{\partial L_1}{\partial S_0} = \frac{\partial L_1}{\partial (S_0 + \Delta S_1)} \cdot \frac{\partial (S_0 + \Delta S_1)}{\partial S_0} = \frac{\partial L_1}{\partial (S_0 + \Delta S_1)} \cdot \left(1 + \frac{\partial \Delta S_1}{\partial S_0}\right) = \frac{\partial L_1}{\partial (S_0 + \Delta S_1)} \cdot \left(1 + \frac{\partial \Delta S_1}{\partial V} \frac{\partial V}{\partial S_0}\right) \quad (10)$$

The terms $\frac{\partial L_1}{\partial (S_0 + \Delta S_1)}$ and $\frac{\partial \Delta S_1}{\partial V}$ have already been computed in Eq. 3. We can separately consider each coordinate x and y of each landmark, i.e. each component of S_0. Thus, for $\frac{\partial V}{\partial S_0}$ and a landmark coordinate x, we have $\frac{\partial V}{\partial x} = \frac{\partial V}{\partial x_p}$ And we can compute $\frac{\partial V}{\partial x_p}$ for each position (q, p) of V by deriving Eq. 5:

$$\frac{\partial V_{qp}}{\partial x_p} = \sum_{n=0}^{H-1} \sum_{m=0}^{W-1} I_{nm} \max\left(0, 1 - \left|y_q - n\right|\right) \begin{cases} 0, & \left|m - x_p\right| \geq 1; \\ 1, & m \geq x_p; \\ -1, & m < x_p. \end{cases} \quad (11)$$

When applying the bilinear interpolation, we can consider only the four neighbouring pixels of (y_q, x_p), therefore the above equation can be simplified to:

$$\frac{\partial V_{qp}}{\partial x_p} = -I_{\lfloor y_q \rfloor \lfloor x_p \rfloor} y_d + I_{\lfloor y_q \rfloor \lceil x_p \rceil} y_d - I_{\lceil y_q \rceil \lfloor x_p \rfloor} y_u + I_{\lceil y_q \rceil \lceil x_p \rceil} y_u \,, \quad (12)$$

where

$$y_d = 1 - \left(y_q - \left\lfloor y_q \right\rfloor\right) \quad (13)$$

$$y_u = 1 - \left(\left\lceil y_q \right\rceil - y_q\right) \,, \quad (14)$$

and $\lfloor\ \rfloor$ and $\lceil\ \rceil$ are the floor and ceiling function respectively. A similar simplification can be applied to Eq. 9.

Hence, we obtain the gradient through the crop location on x coordinate $\frac{\partial V_{qp}}{\partial x_p}$, and analogously for the y coordinate:

$$\frac{\partial V_{qp}}{\partial S_0} = \left(\frac{\partial V_{qp}}{\partial x_p^0}, \frac{\partial V_{qp}}{\partial y_q^0}, ..., \frac{\partial V_{qp}}{\partial x_p^k}, \frac{\partial V_{qp}}{\partial y_q^k}, ...\right), \quad (15)$$

where k indicates the index of each landmark. Now back-propagating gradient of Eq. 10 further until the feature map I gives:

$$\nabla_{crop\_location} = \frac{\partial L_1}{\partial S_0} \cdot \frac{\partial S_0}{\partial I}, \quad (16)$$

where $\frac{\partial S_0}{\partial I}$ can be obtained by standard gradient back-propagation through the main CNN.

5.4. Summary

Consider the standard gradient back-propagated through the baseline network (black arrow in Fig. 7):

$$\nabla_{standard} = \frac{\partial L_0}{\partial I} = \frac{\partial L_0}{\partial S_0} \cdot \frac{\partial S_0}{\partial I}. \quad (17)$$

$\frac{\partial L_0}{\partial S_0}$ is calculated through deriving the loss function, and $\frac{\partial S_0}{\partial I}$ can be obtained by standard gradient back-propagation through the baseline network.

The overall gradient arriving at I through back-propagation is the sum of the three gradients described before:

$$\nabla I = \nabla_{low\_level} + \nabla_{crop\_location} + \nabla_{standard}. \quad (18)$$

6. Experiments

6.1. Datasets

300W dataset: 300W dataset (Sagonas et al., 2013) involves five facial landmark datasets: HELEN (Le et al., 2012), LFPW (Belhumeur et al., 2013), AFW (Zhu and Ramanan, 2012), XM2VTS (Messer et al., 1999) and IBUG. We follow Ren et al. (2014) to use the training set of 3148 images which includes the entire AFW dataset, HELEN training sets and LFPW training sets. The test set of 689 images in total is divided into (i) common subset and (ii) challenging subset. (i) consists of 554 images from the test set of LFPW and HELEN and (ii) consists of 135 images from the IBUG dataset.

300VW dataset: 300VW (Shen et al., 2015) is a video-based facial landmark detection dataset which is annotated in the same manner as 300W. It provides 114 videos in total including 64 videos for validation. The test subset is further categorized into 3 categories based on the level of unconstrained conditions.

AFLW dataset: AFLW (Koestinger et al., 2011) is a large-scale dataset which contains 24386 faces with large pose variations of +/- 90 degree in yaw. All of the images in the dataset are annotated with up to 21 points depending on the landmark visibility. We adopt two protocols from Zhu et al. (2016a). In the *AFLW-Full* protocol, the entire dataset is split into 20,000 images for training and 4,386 images for test. In *AFLW-Frontal* protocol, a subset of 1,314 frontal faces are selected from the entire 4,386 images for frontal evaluation. Note that the landmark format is changed to 19 points excluding the landmarks on both ears in these two protocols.

COFW dataset: COFW (Burgos-Artizzu et al., 2013) is one of the first datasets that aims at benchmarking the performance of facial landmark detection under partial occlusion. It includes 1345 face images for training and 507 face images for testing. All the faces in this dataset are annotated with 29 landmarks.

6.2. Evaluation Metrics

We evaluate our method by measuring the Normalized Mean Error (NME) between our model prediction and the ground truth. On 300W and 300VW datasets, the errors are normalized by the inter-ocular distance (eye-corners or pupils) as in most of the recent comparisons. On the AFLW dataset, due to the large pose variations, we use face size to normalize our mean errors as in Lv et al. (2017). Additionally, the Cumulative Error Distribution (CED) curve is used for evaluation.

6.3. Comparison with State-of-the-art Methods

Results on 300W: A comparison of our methods with other facial landmark detection algorithms is presented in Table 1. In Dong et al. (2018b), we note that the CRMs (Reg+SBR)

Method	Common	Challenge	Full
Inter-Pupil Distance NME (%)			
ESR (Cao et al., 2014)	5.28	17.00	7.58
SDM (Xiong and De la Torre, 2013)	5.57	15.40	7.52
LBF (Ren et al., 2014)	4.95	11.98	6.32
TCDCN (Zhang et al., 2014b)	4.80	8.60	5.54
CFSS (Zhu et al., 2015)	4.73	9.98	5.76
MDM (Trigeorgis et al., 2016)	4.83	10.14	5.88
Lv et al. (2017)	4.36	**7.56**	**4.99**
AAN (Yue et al., 2018)	4.38	9.44	5.39
ECT (Zhang et al., 2018)	4.66	7.96	5.31
DSRN (Miao et al., 2018)	**4.12**	9.68	5.21
ResNet18* (baseline)	6.37	11.32	7.34
ResNet18-FFLD	4.41	8.14	5.14
ResNet50 (baseline)	6.02	10.65	6.94
ResNet50-FFLD	**4.25**	**7.85**	**4.92**
Inter-Eye Corner Distance NME (%)			
PCD-DCNN (Kumar and Chellappa, 2018)	3.67	7.62	4.44
JDR (Zhu et al., 2019a)	3.68	7.16	4.36
SAN (Dong et al., 2018a)	3.34	**6.60**	**3.98**
Reg + SBR (Dong et al., 2018b)	7.93	15.98	9.46
CPM + SBR (Dong et al., 2018b)	**3.28**	7.58	4.10
ODN (Zhu et al., 2019b)	3.56	6.67	4.17
ADN Sadiq et al. (2019)	3.50	6.60	4.14
ResNet18 (baseline)	4.38	7.46	4.98
ResNet18-FFLD	3.18	5.64	3.66
ResNet50 (baseline)	4.12	7.05	4.73
ResNet50-FFLD	**3.06**	**5.44**	**3.50**

Table 1: Comparison of Normalized Mean Error (NME) on 300W dataset of ResNet with our Fine-grained Facial Landmark Detection (ResNet18/50-FFLD) framework and other approaches. * Note that all ResNet18/50 used here are simplified version with only 25% channels compared to the original version.

has inferior performance compared to the HRMs (CPM+SBR). By integrating our Fine-grained Facial Landmark Detection (FFLD) framework, our CRM based model has a comparable performance to state-of-the-art HRMs (Kumar and Chellappa, 2018; Dong et al., 2018b,a; Tai et al., 2019). Refined prediction has been significantly improved by nearly 25% compared to the baseline output. We show our CED curve compared to 3DDFA (Zhu et al., 2016b), DRMF (Asthana et al., 2013), CFSS (Zhu et al., 2015) and TCDCN (Zhang et al., 2014b) in Fig. 9 (b). We observed that our coarse-to-fine FFLD framework (gray) is able to provide precise fine-grained correction to the coarse prediction given by the baseline network (pink). Several qualitative results are presented in Fig. 10. Specifically, by using our proposed Align Loss, we found that the landmarks are more aligned on the boundaries of the facial components. An additional visual comparison is shown in Fig. 15 and Fig. 16.

Results on AFLW: We show the performance comparison on the AFLW dataset in Table 2. Compared to our baseline, the precision of ResNet18-FFLD is significantly improved by a large margin of 25%. We compare our methods with LBF (Ren et al., 2014), ERT (Kazemi and Sullivan, 2014), CFSS (Zhu et al., 2015), SDM (Xiong and De la Torre, 2013), CCL (Zhu et al., 2016a), DAC-CSR (Feng et al., 2017) in cumulative error distribution curve shown in Fig. 9 (a). We visually compare the

Method	AFLW-Full	AFLW-Front
SDM	4.05	2.94
ERT	4.35	2.75
LBF	4.25	2.74
CFSS	3.92	2.68
CCL (Zhu et al., 2016a)	2.72	2.17
DAC-CSR (Feng et al., 2017)	2.27	1.81
Reg + SBR	4.77	-
CPM + SBR	2.14	-
JDR	1.97	1.69
SAN	1.91	1.85
DSRN	1.86	-
ODN	**1.63**	**1.38**
ResNet18*	2.30	1.99
ResNet18-FFLD	1.75	1.52
ResNet50	2.13	1.86
ResNet50-FFLD	**1.62**	**1.42**

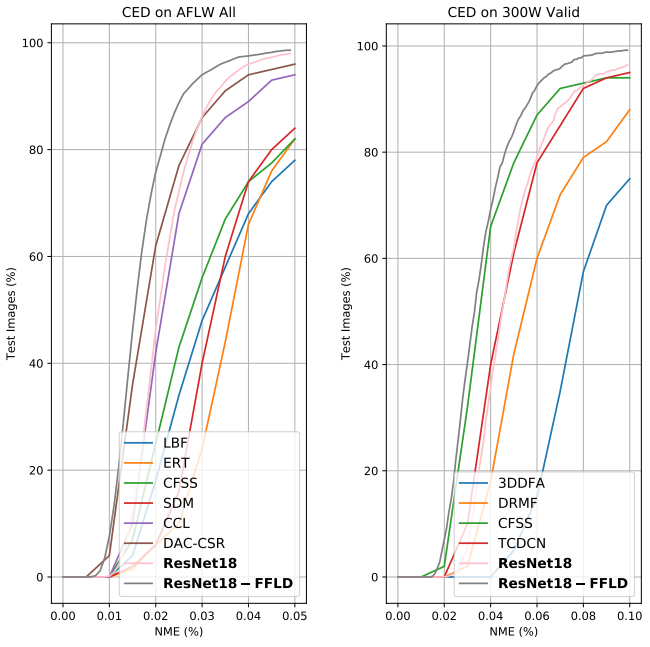
Table 2: Comparison of face size NME (%) on AFLW dataset of ResNet18/50-FFLD and other approaches. * Note that all ResNet18/50 used here are simplified version with only 25% channels compared to the original version.

Method	NME
ESR	11.20
CCR	7.03
DAC-CSR	6.03
ECT	5.98
ODN (Zhu et al., 2019b)	**5.30**
ResNet18*	6.84
ResNet18-FFLD	5.45
ResNet50	6.53
ResNet50-FFLD	**5.32**

Table 3: Comparison of face size NME (%) on COFW dataset of ResNet18/50-FFLD and other approaches. * Note that all ResNet18/50 used here are simplified version with only 25% channels compared to the original version.

Method	Cat. 1	Cat. 2	Cat. 3
TCDCN	7.66	6.77	15.00
CFSS	7.68	6.42	13.70
HG (Newell et al., 2016)	5.44	4.71	7.92
TSTN (Liu et al., 2018)	5.36	4.51	12.80
DSRN	5.33	4.92	8.85
FHR (Tai et al., 2019)	5.07	**4.34**	7.36
ADN (Sadiq et al., 2019)	**4.69**	**4.34**	**6.72**
ResNet18	5.86	5.12	9.14
ResNet18-FFLD	**4.85**	**4.24**	**7.62**

Table 4: Comparison of NME (%) on the 300VW dataset of ResNet18-FFLD and other approaches.



(a) AFLW (b) 300W

Fig. 9: The Cumulative Error Distribution (CED) curves of our method on AFLW and 300W datasets.

prediction between our approach and the baseline in Fig. 17.

Results on COFW: The comparison is shown in Table 3. Our method achieves similar performance compared to a recent method ODN (Zhu et al., 2019b). Note that our method is not specifically designed for occluded images (local boundary information may be obscured by occlusions). We prove here that our method is still robust to the occlusion, though highly dependent on the local detail information.

Results on 300VW: A comparison of ResNet18-FFLD with other methods on the 300VW dataset is shown in Table 4. The 300VW dataset contains frames with large poses. We apply facial landmark detection in a frame-by-frame manner without exploiting any temporal information. Compared to the baseline network, the performance is improved by more than 20% with our framework. Compared to other state-of-the-art methods, our method achieves superior or similar performance on Cat. 1 and Cat. 2. However, the faces are in rather low resolution in Cat. 3. Therefore, our method does not outperform ADN (Sadiq et al., 2019) on the Cat. 3. Further discussions on this issue will be presented in Sect. 6.5 Failure Cases.

6.4. Ablation Studies

Multi-stage & multi-loss comparison: The improvement of our method originates from two aspects: (1) the use of CropNet (multi-stage refinement) and (2) the use of different loss functions in different stages. In Table 5, we compare the results of different loss functions and different number of refinement stages on the 300W dataset based on the ResNet-18 baseline.

With more refinement stages, the precision is progressively improved. Specifically, when using 3 refinement stages, we test different combinations of the loss functions. Compared to using the same loss function for all stages ($L2$, $L1$ or AL loss), using a combination of different loss functions can additionally improve the performance. By assigning the loss functions that are more sensitive to the small errors in the last stages, the overall performance is further improved. Although this improvement is numerically incremental, it is nonetheless critical for many aesthetic rendering applications. We visually illustrate this improvement in Fig. 10 and Fig. 15.

Gradient backpropagation: To demonstrate that the gradient ∇_{low_level} and $\nabla_{crop\_location}$ do help to improve the refinement. We show the performance by disabling gradient back-propagation on the baseline network feature maps in Table 6. We found that the NME on 300W is further improved from 3.78% to 3.66% by allowing both ∇_{low_level} and $\nabla_{crop\_location}$ to be back-propagated on the low-level feature maps of the baseline network. Though both gradient back-propagation achieves incremental improvement, we observe that ∇_{low_level} plays a more important role compared to $\nabla_{crop\_location}$.

Effectiveness of our method on small errors: We will now focus on the small errors when performing FFLD. To further prove the effectiveness of using different losses and back propagation strategies on small errors, we show a landmark-wise CED in Fig. 11. In this figure, we focus on the landmark-wise NME from 1.0 to 3.0. We also sample the difference of the models at NME=2.75. We observe that by using a unique $L2$ loss for all refinement stages, the improvement is limited when the third refinement stage is added (green & red). When multi-

Fig. 10: Qualitative results of our approach on the 300W dataset. The prediction by the baseline network ResNet18 (the 1st row). The prediction by ResNet18 with our Fine-grained Facial Landmark Detection (ResNet18-FFLD) framework **without Align Loss** (the 2nd row). The prediction by ResNet18 with FFLD framework **with Align Loss** (the 3rd row). To better visualize the small error, we provide the zoomed image aside. More examples are provided in Fig. 15.

Method	Baseline (w/o Ref)	2 Stages Ref		3 Stages Ref				
Loss	Norm *L*2	*L*2/*L*2	*L*2/*L*1	*L*2/*L*2/*L*2	*L*1/*L*1/*L*1	AL/AL/AL	*L*2/*L*1/*L*1	*L*2/*L*1/AL
NME	4.98	3.85	3.77	3.81	3.76	3.95	3.72	**3.66**

Table 5: Multi-loss ablation study on 300W with ResNet-18 as baseline network. Ref - Refinement. AL - Align Loss.

∇_{low_level}	$\nabla_{crop\_location}$	NME
✗	✗	3.78
✗	✓	3.74
✓	✗	3.69
✓	✓	**3.66**

Table 6: Gradient back-propagation ablation study on 300W with ResNet-18.

loss scheme is applied (blue), the performance can be improved by a large margin. We also observe that by enabling the ∇_{low_level} and $\nabla_{crop\_location}$ on the low-level feature maps, the performance on small errors can be further improved.

IOD normalized *L*2 **loss:** We also compared the results of our baseline ResNet network between using IOD normalized *L*2 loss and standard *L*2 loss. By training with IOD Normalized *L*2 loss, the inter eye-corner distance NME on 300W is improved from 5.14% to 4.98%.

6.5. Discussions

Run time and model size: Without any speed optimization such as MobileNet blocks (Sandler et al., 2018), our 3-stage ResNet18-FFLD model runs at 130 fps on a NVIDIA 1080Ti GPU. Our model contains 1.46M parameters, including the baseline network. With an input size of 224×224 and a batch size of 64, our networks require less than 2GB GPU memory during training compared to the heat map regression model in Wu et al. (2018) which require 4 NVIDIA Titan X GPUs to train with a batch size of 8.

Robustness: While being highly dependent on the boundary information on the low-level feature map, we found that our CropNet is robust to partial occlusions (see Fig. 12). In Fig. 13, we show a histogram of the CropNet output ΔS in each stage. We found that the ΔS forms a stable distribution without long tail, which proves the robustness of our model. We think that ΔS is always regularized by the overall shape because the ΔS for each landmark are predicted by the same FC layer. Previous experimental results on COFW dataset could also support this conclusion.

Failure cases: We show three worst cases of our detection. All of them are on rather low resolution test images (see Fig. 14). In this case, CropNet has difficulties in capturing enough meaningful details (e.g. boundaries) from the indistinct low-level feature maps.

Connection with cascaded CNNs: Most of the cascaded patch-based structures Chen et al. (2017) depend on the image patches, which require the refinement stage to learn from RGB information. Our approach is the first to focus on feature map patches. Our patch-based method stands out for the following reasons: (a) The gradient from the refinement stage can be back-propagated directly on the feature maps of the main network. Refinement networks are learned end-to-end, jointly with the main network. (b) It is easier to learn with boundary information from the feature map patches than RGB information from the image patches. (c) With identical patch support, the spatial dimension of feature map patches is much smaller than image patches, which is computationally less expensive.

Connection with HRMs: Both our skip connection and spatial robust loss function are inspired by the popular heat map regression models. Our method and HRMs can be both trained end-to-end. Moreover, our approach is more memory-efficient than existing heat map regression models thanks to the smaller

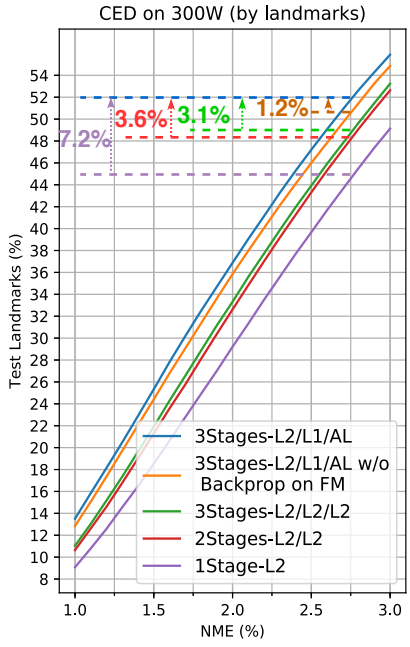


Fig. 11: Landmark-wise CED focused on small errors. AL-Align Loss. w/o Backprop on FM: neither ∇_{low_level} nor $\nabla_{crop\_location}$ is back-propagated on the low-level feature maps. All models are based on the ResNet-18 baseline and tested on 300W.

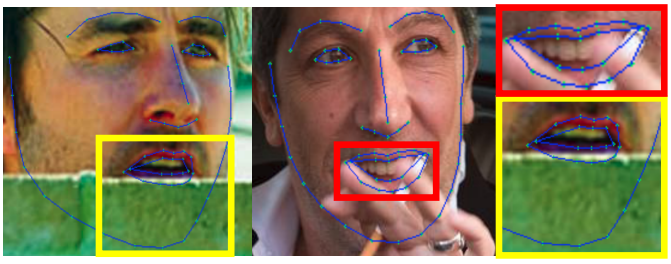


Fig. 12: Examples of our detection on partially occluded images. We can observe that our detection is robust to the occlusions on the face images. That means that even when the boundary information is not given, CropNet still provides a reasonable shape as output.

spatial dimension in skip connections.

6.6. Implementation Details

For CropNet, we propose a relatively simple structure: one batch normalization layer, one ResNet block, one max-pooling layer and one FC layer (see Table 7).

For the baseline CNN, we used ResNet18/50 (He et al., 2016) but reduced 75% of its feature map channels in each layer. One important challenge for facial landmark detection is to correctly detect the facial landmarks on extreme poses (Sagonas et al., 2013; Zafeiriou et al., 2017). In this context, Feng et al. (2018) argued that this issue is due to the imbalanced data distribution. To balance the examples in different poses and construct a pose-robust model for landmark refinement on 300W and 300VW, we followed the training strategy in Bulat and Tzimiropoulos (2017b). We first pre-trained the model on a large synthetic

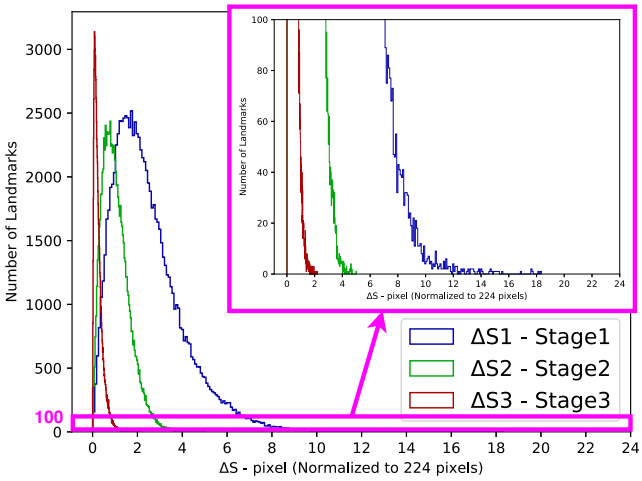


Fig. 13: Histogram of ΔS in each refinement stage (by landmark). The ΔS of each CropNet stage forms a stable distribution without long tail.

Fig. 14: Failure cases on low resolution images.

dataset 300W-LP (Zhu et al., 2016b) (LP means Large Pose) with a learning rate of 0.0002 for 80 epochs. 300W-LP expands the 300W dataset by synthesizing large-pose face appearances with a 3D face model and rendering them in different poses (no extra faces). However, the 2D annotation in 300W-LP is not compatible with the original 300W annotation. We then trained our model on 300W dataset for another 350 epochs. The learning rate starts from 0.0001 with a decay of 0.3 for each 50 epochs. On AFLW, we used the PDB strategy from Feng et al. (2018) to overcome the imbalanced data distribution problem. We trained our model with a learning rate of 0.0001 for 56 epochs. The learning rate is decayed by 0.3 every 8 epochs.

Afterwards, we trained our 3-stage ResNet-FFLD with three different losses for 400 epochs. The learning rate is initialized to 0.0005 and decayed by 0.3 for each 80 epochs. We initialized the weights of FC layers in CropNets to zero. For the Align Loss, the initial Gaussian kernel size of the convolution kernel is 3 and the sigma is 1. In order to achieve extreme precision, this operation is removed once the loss stops going down.

All experiments are conducted with PyTorch. We used a batch size of 64, Adam as optimizer and 0.0005 as weight decay for all of the training. We further applied data augmentation of +/- 20% in scale, +/- 10% in vertical/horizontal translation and +/- 20% in rotation.

Layer	(In_channels, Out_channels, Stride)
Batch Norm	$(4 \times N, 4 \times N, 1)$
ResNet Block	$(4 \times N, 128, 1)$
Max-pooling	(128, 128, 8)
FC	$(128, 2 \times N, -)$

Table 7: The structure of our CropNet. The right column shows the parameters for each layer/block including input channels, output channels and stride. N denotes the number of facial landmarks, which can vary for different datasets.

7. Conclusions and Future Work

We presented an novel effective end-to-end framework for Fine-grained Facial Landmark Detection based on coordinate regression deep neural network models. We showed that low-level feature maps contain important contour information, that is useful for refining the landmark positions. By establishing skip connections, the localization accuracy of a coordinate regression model can be significantly improved and achieves comparable performance to the state-of-the-art heat map regression models. In addition, training different refinement stages with different loss functions, including the proposed Align Loss which forces the landmark to learn extreme accurate prediction, can further increase the localization precision.

An interesting future direction of this work is to delve deeper into the differences between CRMs and HRMs. In practice, we find that it is much easier to train an HRM than a CRM. Though HRMs are less efficient during inference, HRMs require much less number of epochs during training. We assume that the output representation of HRMs is more intuitive than CRMs for CNNs. Future work can be focused on how to ease the training process of the CRMs, through merging the output representation of HRMs with CRMs.

Acknowledgments

This work is funded by the Auvergne Regional Council and the European funds of regional development (FEDER). The computation resource isupported by Msocentre Clermont Auvergne. We would like to thank Nvidia for GPU donation.

References

Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M., 2013. Robust discriminative response map fitting with constrained local models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3444–3451.

Belagiannis, V., Rupprecht, C., Carneiro, G., Navab, N., 2015. Robust optimization for deep regression, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2830–2838.

Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N., 2013. Localizing parts of faces using a consensus of exemplars. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 2930–2940.

Bulat, A., Tzimiropoulos, G., 2017a. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3706–3714.

Bulat, A., Tzimiropoulos, G., 2017b. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks), in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1021–1030.

Burgos-Artizzu, X.P., Perona, P., Dollár, P., 2013. Robust face landmark estimation under occlusion, in: The IEEE International Conference on Computer Vision (ICCV).

Cao, X., Wei, Y., Wen, F., Sun, J., 2014. Face alignment by explicit shape regression. International Journal of Computer Vision 107, 177–190.

Chang, C.H., Chou, C.N., Chang, E.Y., 2017. Clkn: Cascaded lucas-kanade networks for image alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2213–2221.

Chen, X., Zhou, E., Mo, Y., Liu, J., Cao, Z., 2017. Delving deep into coarse-to-fine framework for facial landmark localization., in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2088–2095.

Chen, Y., Shen, C., Chen, H., Wei, X.S., Liu, L., Yang, J., 2019. Adversarial learning of structure-aware fully convolutional networks for landmark localization. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Ding, C., Tao, D., 2016. A comprehensive survey on pose-invariant face recognition. ACM Transactions on Intelligent Systems and Technology 7, 37:1–37:42.

Dong, X., Yan, Y., Ouyang, W., Yang, Y., 2018a. Style aggregated network for facial landmark detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 379–388.

Dong, X., Yu, S.I., Weng, X., Wei, S.E., Yang, Y., Sheikh, Y., 2018b. 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, pp. 360–368.

Duffner, S., Garcia, C., 2005. A connexionist approach for robust and precise facial feature detection in complex scenes, in: Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis., pp. 316–321.

Fan, H., Zhou, E., 2016. Approaching human level facial landmark localization by deep learning. Image and Vision Computing 47, 27–35.

Feng, Z.H., Kittler, J., Awais, M., Huber, P., Wu, X.J., 2018. Wing loss for robust facial landmark localisation with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2235–2245.

Feng, Z.H., Kittler, J., Christmas, W., Huber, P., Wu, X.J., 2017. Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3681–3690.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017a. Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

He, Z., Kan, M., Zhang, J., Chen, X., Shan, S., 2017b. A fully end-to-end cascaded cnn for facial landmark detection, in: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition, pp. 200–207.

Honari, S., Molchanov, P., Tyree, S., Vincent, P., Pal, C., Kautz, J., 2018. Improving landmark localization with semi-supervised learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1546–1553.

Jackson, A.S., Bulat, A., Argyriou, V., Tzimiropoulos, G., 2017. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. Proceedings of the IEEE International Conference on Computer Vision .

Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks, in: Advances in Neural Information Processing Systems, pp. 2017–2025.

Kazemi, V., Sullivan, J., 2014. One millisecond face alignment with an ensemble of regression trees, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867–1874.

Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H., 2011. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization, in: The IEEE International Conference on Computer Vision (ICCV) Workshops, pp. 2144–2151.

Kowalski, M., Naruniec, J., Trzcinski, T., 2017. Deep alignment network: A convolutional neural network for robust face alignment, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, p. 6.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with

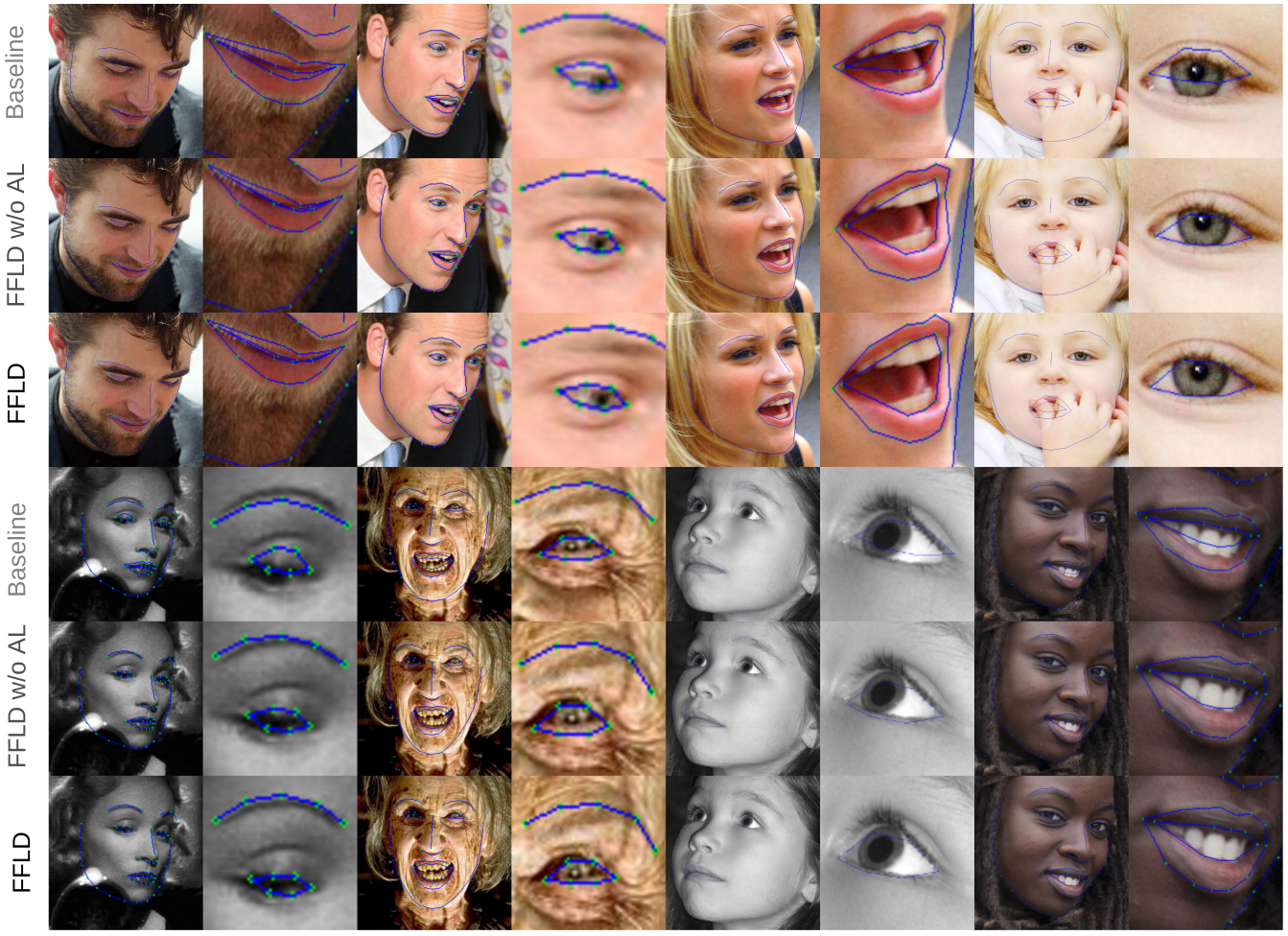


Fig. 15: Qualitative results of our approach on the 300W dataset. The prediction by the baseline network ResNet18 (the 1st row). The prediction by ResNet18 with our Fine-grained Facial Landmark Detection (ResNet18-FFLD) framework **without Align Loss** (the 2nd row). The prediction by ResNet18 with FFLD framework **with Align Loss** (the 3rd row). To better visualize the small error, we provide the zoomed image aside.

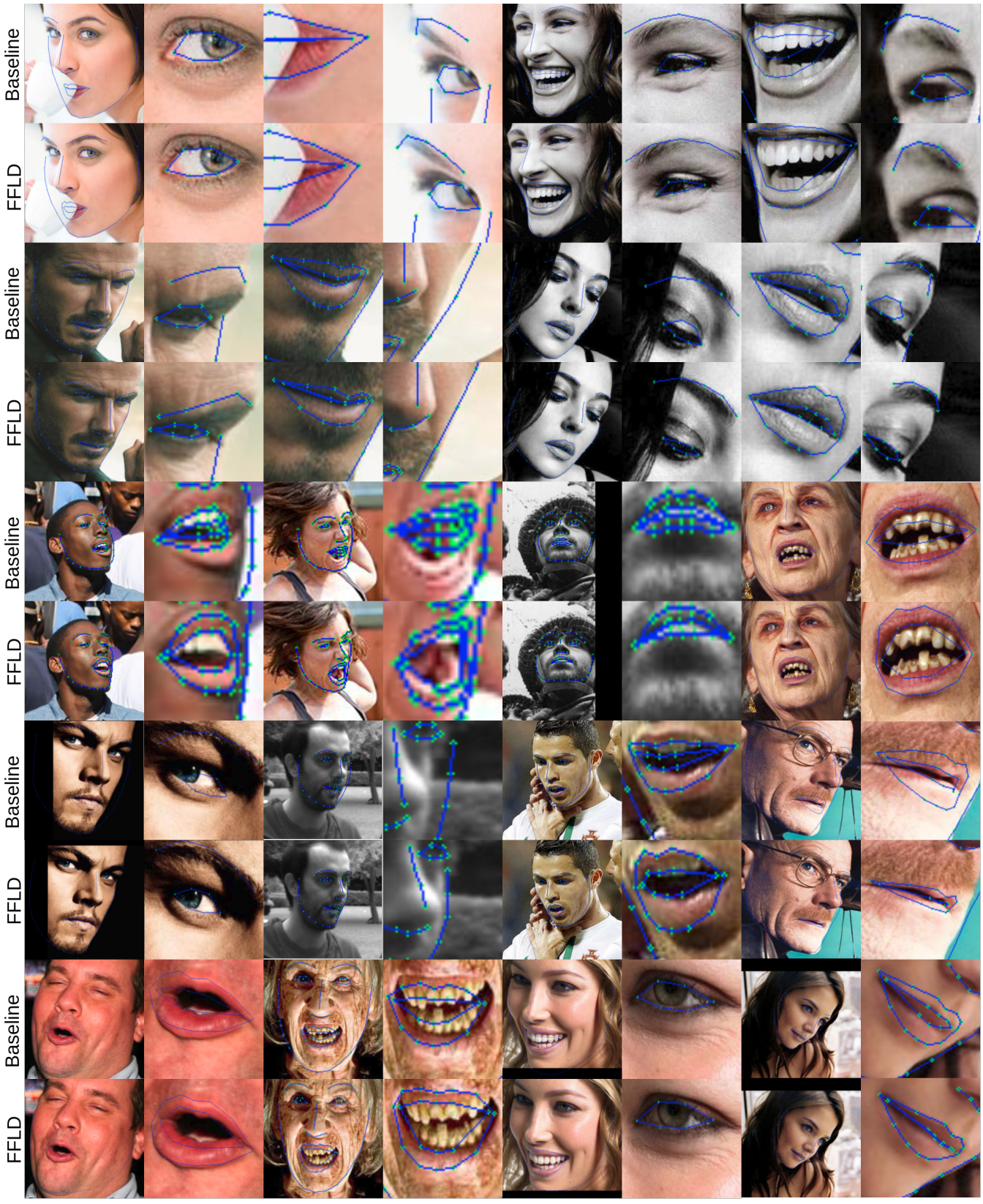


Fig. 16: Qualitative results of our approach on the 300W dataset. The prediction by the baseline network ResNet18 (first row). The prediction by ResNet18 with our Fine-grained Facial Landmark Detection (ResNet18-FFLD) framework (second row). To better visualize the small error, we provide the zoomed image aside.

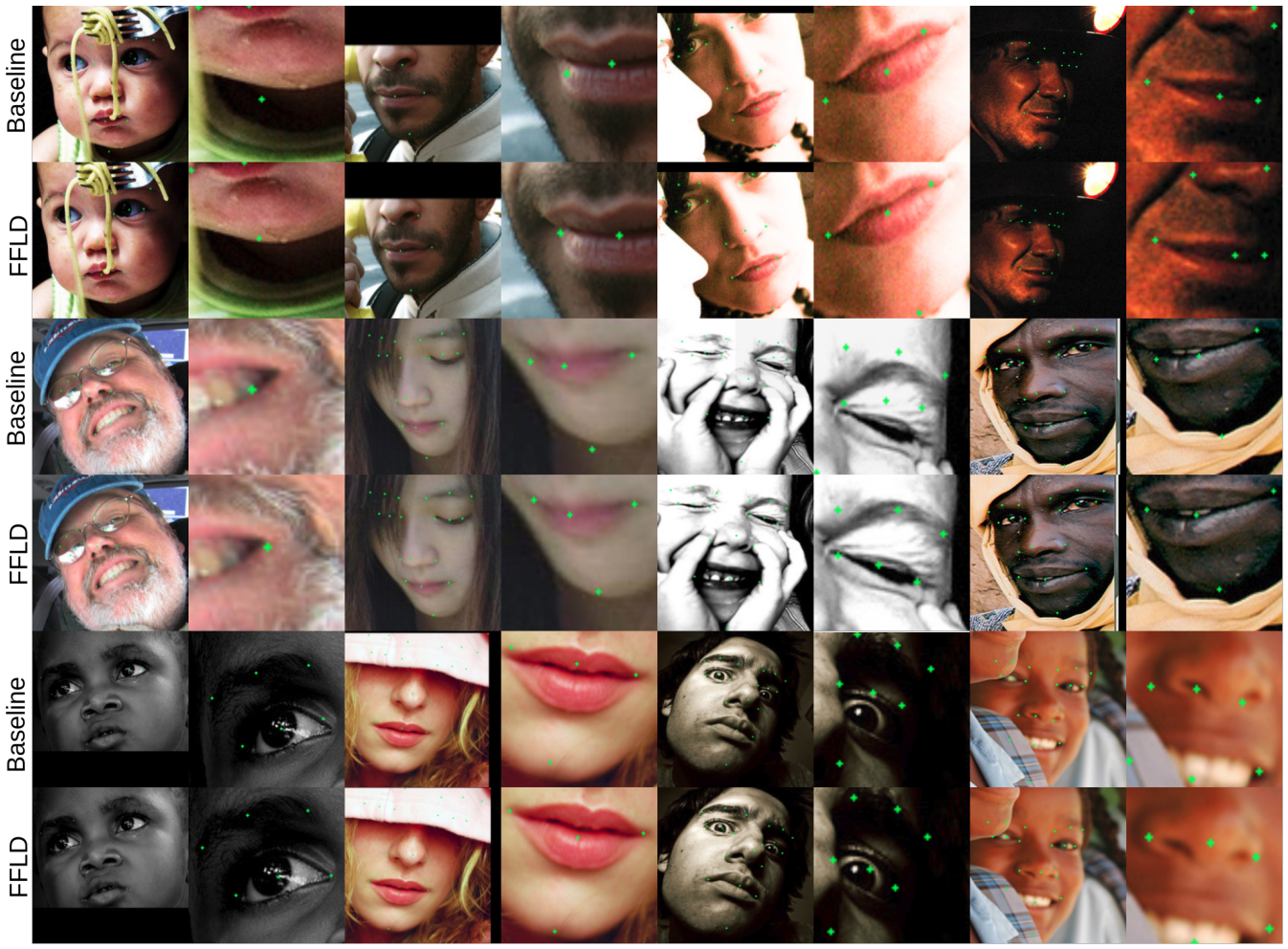


Fig. 17: Qualitative results of our approach on the AFLW dataset. The prediction by the baseline network ResNet18 (first row). The prediction by ResNet18 with our Fine-grained Facial Landmark Detection (ResNet18-FFLD) framework (second row). To better visualize the small error, we provide the zoomed image aside.

deep convolutional neural networks, in: Advances in Neural Information Processing Systems, pp. 1097–1105.

Kumar, A., Chellappa, R., 2018. Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 430–439.

Lai, H., Xiao, S., Pan, Y., Cui, Z., Feng, J., Xu, C., Yin, J., Yan, S., 2018. Deep recurrent regression for facial landmark detection. IEEE Transactions on Circuits and Systems for Video Technology 28, 1144–1157.

Lathuilière, S., Mesejo, P., Alameda-Pineda, X., Horaud, R., 2018. Deepgum: Learning deep robust regression with a gaussian-uniform mixture model, in: Proceedings of the European Conference on Computer Vision, pp. 205–221.

Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S., 2012. Interactive facial feature localization, in: Proceedings of the European Conference on Computer Vision, pp. 679–692.

Liu, H., Lu, J., Feng, J., Zhou, J., 2018. Two-stream transformer networks for video-based face alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, 2546–2554.

Lv, J., Shao, X., Xing, J., Cheng, C., Zhou, X., 2017. 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, pp. 3317–3326.

Martinez, B., Valstar, M.F., Jiang, B., Pantic, M., 2017. Automatic analysis of facial actions: A survey. IEEE Transactions on Affective Computing .

Merget, D., Rock, M., Rigoll, G., 2018. Robust facial landmark detection via a fully-convolutional local-global context network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 781–790.

Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G., 1999. Xm2vtsdb: The extended m2vts database, in: International Conference on Audio and Video-based Biometric Person Authentication, pp. 965–966.

Miao, X., Zhen, X., Liu, X., Deng, C., Athitsos, V., Huang, H., 2018. Direct shape regression networks for end-to-end face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5040–5047.

Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation, in: Proceedings of the European Conference on Computer Vision, pp. 483–499.

Nibali, A., He, Z., Morgan, S., Prendergast, L., 2018. Numerical coordinate regression with convolutional neural networks. arXiv preprint arXiv:1801.07372 .

Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R., 2017. An all-in-one convolutional neural network for face analysis, in: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition, pp. 17–24.

Ren, S., Cao, X., Wei, Y., Sun, J., 2014. Face alignment at 3000 fps via regressing local binary features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1685–1692.

Sadiq, M., Shi, D., Guo, M., Cheng, X., 2019. Facial landmark detection via attention-adaptive deep network. IEEE Access 7, 181041–181050.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M., 2013. 300 faces in-the-wild challenge: The first facial landmark localization challenge, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 397–403.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4517.

Shen, J., Zafeiriou, S., Chrysos, G.G., Kossaifi, J., Tzimiropoulos, G., Pantic, M., 2015. The first facial landmark tracking in-the-wild challenge: Benchmark and results, in: The IEEE International Conference on Computer Vision (ICCV) Workshops, pp. 50–58.

Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y., 2018. Integral human pose regression, in: Proceedings of the European Conference on Computer Vision, pp. 529–545.

Sun, Y., Wang, X., Tang, X., 2013. Deep convolutional network cascade for facial point detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3476–3483.

Tai, Y., Liang, Y., Liu, X., Duan, L., Li, J., Wang, C., Huang, F., Chen, Y., 2019. Towards highly accurate and stable face alignment for high-resolution videos, in: Proceedings of the Conference on Artificial Intelligence (AAAI).

Tang, Z., Peng, X., Geng, S., Wu, L., Zhang, S., Metaxas, D., 2018. Quantized densely connected u-nets for efficient landmark localization, in: Proceedings of the European Conference on Computer Vision.

Trigeorgis, G., Snape, P., Nicolaou, M.A., Antonakos, E., Zafeiriou, S., 2016. Mnemonic descent method: A recurrent process applied for end-to-end face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4177–4187.

Valle, R., Buenaposada, J.M., Valdes, A., Baumela, L., 2018. A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment, in: Proceedings of the European Conference on Computer Vision.

Wang, Z., Liu, S., Hu, W., Tong, R., Yang, X., Zhang, J.J., 2017. Face alignment refinement via exploiting low-rank property and temporal stability, in: Proceedings of the 30th International Conference on Computer Animation and Social Agents (CASA).

Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y., 2016. Convolutional pose machines, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4724–4732.

Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q., 2018. Look at boundary: A boundary-aware face alignment algorithm, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2129–2136.

Wu, Y., Ji, Q., 2017. Facial landmark detection: A literature survey. International Journal of Computer Vision , 1–28.

Xiong, X., De la Torre, F., 2013. Supervised descent method and its applications to face alignment, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 532–539.

Yan, Y., Naturel, X., Chateau, T., Duffner, S., Garcia, C., Blanc, C., 2018. A survey of deep facial landmark detection, in: Reconnaissance des Formes, Image, Apprentissage et Perception RFIAP.

Yu, X., Zhou, F., Chandraker, M., 2016. Deep deformation network for object landmark localization, in: Proceedings of the European Conference on Computer Vision, pp. 52–70.

Yue, L., Miao, X., Wang, P., Zhang, B., Zhen, X., Cao, X., 2018. Attentional alignment networks, in: Proceedings of the British Machine Vision Conference, p. 208.

Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., Shen, J., 2017. The menpo facial landmark localisation challenge: A step towards the solution, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, p. 2.

Zeng, A., Boddeti, V.N., Kitani, K.M., Kanade, T., 2015. Face alignment refinement, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, pp. 162–169.

Zhang, H., Li, Q., Sun, Z., Liu, Y., 2018. Combining data-driven and model-driven methods for robust facial landmark detection. IEEE Transactions on Information Forensics and Security 13, 2409–2422.

Zhang, J., Hu, H., 2018. Exemplar-based cascaded stacked auto-encoder networks for robust face alignment. Computer Vision and Image Understanding 171, 95–103.

Zhang, J., Shan, S., Kan, M., Chen, X., 2014a. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment, in: Proceedings of the European Conference on Computer Vision, Springer. pp. 1–16.

Zhang, Z., Luo, P., Loy, C.C., Tang, X., 2014b. Facial landmark detection by deep multi-task learning, in: Proceedings of the European Conference on Computer Vision, pp. 94–108.

Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q., 2013. Extensive facial landmark localization with coarse-to-fine convolutional network cascade, in: The IEEE International Conference on Computer Vision (ICCV) Workshops, pp. 386–391.

Zhu, M., Shi, D., Gao, J., 2019a. Branched convolutional neural networks incorporated with jacobian deep regression for facial landmark detection. Neural Networks 118, 127–139.

Zhu, M., Shi, D., Zheng, M., Sadiq, M., 2019b. Robust facial landmark detection via occlusion-adaptive deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3486–3493.

Zhu, S., Li, C., Change Loy, C., Tang, X., 2015. Face alignment by coarse-to-fine shape searching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4998–5006.

Zhu, S., Li, C., Loy, C.C., Tang, X., 2016a. Unconstrained face alignment via cascaded compositional learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3409–3417.

Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z., 2016b. Face alignment across large poses: A 3d solution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 146–155.

Zhu, X., Ramanan, D., 2012. Face detection, pose estimation, and landmark lo-

calization in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2879–2886.