Robust Facial Landmark Detection via a Fully-Convolutional Local-Global Context Network

Daniel Merget daniel.merget@tum.de
Matthias Rock matthias.rock@tum.de
Gerhard Rigoll mmk@ei.tum.de
Technical University of Munich

Abstract

While fully-convolutional neural networks are very strong at modeling local features, they fail to aggregate global context due to their constrained receptive field. Modern methods typically address the lack of global context by introducing cascades, pooling, or by fitting a statistical model. In this work, we propose a new approach that introduces global context into a fully-convolutional neural network directly. The key concept is an implicit kernel convolution within the network. The kernel convolution blurs the output of a local-context subnet, which is then refined by a global-context subnet using dilated convolutions. The kernel convolution is crucial for the convergence of the network because it smoothens the gradients and reduces overfitting. In a postprocessing step, a simple PCA-based 2D shape model is fitted to the network output in order to filter outliers. Our experiments demonstrate the effectiveness of our approach, outperforming several state-of-the-art methods in facial landmark detection.

1. Introduction

Facial landmark detection is a well-studied topic in the field of computer vision with many applications such as face analysis, face recognition, or face modeling; see [22] for a review. The high variability of shapes, poses, lighting conditions, and possible occlusions makes it a particulary challenging task even today. In contrast to face recognition, where modern approaches using convolutional neural networks (CNNs) are beyond human-level performance [33], computers are still below par at facial landmark detection [8].

Classical CNN-based architectures start off with a number of convolutional layers that are aggregated by fully-connected layers at the end. Recently, the trend has shifted towards using fully-convolutional networks, dispensing with full connections. Fully-convolutional networks did not only prove successful in facial landmark detection [40], but also in other applications such as image classification [26, 32] or object detection [25]. This is because fully-convolutional architectures come with several advantages that are especially useful for heatmap regression tasks in general and facial landmark detection in particular:

1. They are independent of image resolution
2. They do not depend on proper regions of interest such as bounding boxes from a face detector
3. They can handle both empty and multiple outputs per input (*e.g.*, multiple faces)
4. They can handle cropped or occluded images to a reasonable extent, degrading gracefully
5. They require fewer trainable parameters and consequently have a lower memory footprint

Fully-convolutional networks also come with a huge drawback: They have a limited receptive field and thus lack global context [21]. In other words, global context must be introduced by other measures that do not share the disadvantages of fully-connected architectures [24].

Our method builds on the idea that global context can be integrated in a fully-convolutional network by rigorously extending the effective receptive field of the network. This extension is achieved by a channel-wise convolution with a two-dimensional kernel, or simply *kernel convolution*, followed by a global-context subnet based on dilated convolutions. We will refer to the extended local context as *global context*, but learned patterns are still purely *relative* rather than *absolute*. For example, there is no bias on the face location within the image due to the fully-convolutional design.

The major contributions of this work are the following:

- We incorporate a kernel convolution directly within a neural network.
- We exploit the kernel convolution in order to robustly increase the receptive field (*i.e.*, context) of the network using dilated convolutions.

- We demonstrate the effectiveness of our approach by improving upon the state of the art on 300-W [28] and a cross-data set test on Menpo [41].
- We demonstrate qualitatively that our approach does not depend on prior face detections

2. Related work

Related work can be divided into two general groups: Facial landmark detection-specific methods and heatmap regression methods relying on fully-convolutional architectures. The former methods can be subdivided further into *model-based* and *regression-based* methods.

2.1. Model-based landmark detection

Saragih *et al*. [30] built a constrained local model based on the concept of a regularized mean shift. Another constrained local model based on neural fields (CLNF) was developed by Baltrušaitis *et al*. [4]. More recently, He *et al*. [12] expressed the problem of landmark detection as a non-linear mapping between an input image and a shape, which they modeled via a deep cascade of convolutional subnets. Specifically addressing profile faces, a 3D dense face alignment (3DDFA) was suggested by Zhu *et al*. [45] who used a CNN to update the parameters of a 3D morphable model. Similarly, Güler *et al*. [11] built a CNN to regress a statistical shape model by computing a bijective mapping from a 3D template mesh to the 2D image.

2.2. Regression-based landmark detection

Zhang *et al*. [43] trained a task-contrained CNN to a set of features trained from multiple tasks (TCDCN). Sun *et al*. [31] suggested a coarse-to-fine shape regression method and, similarly, Zhu *et al*. [44] performed a cascaded coarse-to-fine shape searching (CFSS). Xiong and De la Torre [37] developed the supervised descent method (SDM) which iteratively finds the shape of the face by minimizing an L2 loss. It is also possible to linearly regress the features extracted by a coarse-to-fine autoencoder (CFAN) as demonstrated by Zhang *et al*. [42]. Tzimiropoulos [35] developed the method of project out cascaded regression (PO-CR) which is another cascaded regression technique. Xiao *et al*. [36] built a recurrent 3D regressor network that iteratively refines the output of a (regressional) feature network.

Regressing heatmaps from local patch experts, Asthana *et al*. [1] propose a discriminative response map fitting approach (DRMF). Finally, Zadeh *et al*. [40] used a constrained local model supported by convolutional experts to regress facial landmarks from heatmaps (CE-CLM). Both Asthana *et al*. [1] and Zadeh *et al*. [40] train a set of local patch experts; global context is later induced by a point distribution model. That is, global context is only considered by the models, not by the patch responses.

2.3. Fully-convolutional heatmap regression

Heatmap regression in the domain of deep learning is most commonly used for dense pixel-wise classification problems such as foreground-background segmentation [2] or object segmentation [3, 23]. Recently, however, heatmap regression has been applied successfully to sparse problems such as human pose estimation [6, 24, 34]. While all three of these works found that using a kernel convolution is beneficial, they did not incorporate this convolution into the network. Their networks thus have to waste resources on learning how to reconstruct the kernel in the output. Pfister *et al*. [24] furthermore noticed the lack of global context in their fully-convolutional network and consequently increased the receptive field by using larger filters (*e.g*., 15×15) for the convolutional layers in a separate *spatial fusion* branch. The problem with larger filters is that they are overfitting very easily and the increase in context is very limited.

Context aggregation of heatmaps with the help of dilated convolutions was first proposed by Yu and Koltun [39] in the setting of object segmentation. They do not use any kernel, which works well in their application because objects are large compared to facial landmarks.

3. Local-global context network

We pursue a heatmap regression approach similar to Zadeh *et al*. [40]. The major advantage of our approach is that the *network* takes into account global context, which is then *refined* by a point distribution model during post-processing. Furthermore, our network is more robust since it does not rely on region proposals, which constitute an additional source of errors. Both Zadeh *et al*. [40] and Pfister *et al*. [24] place a normal distribution on the ground truth labels. In contrast to their approaches, our kernel convolution is applied implicitly, which allows the network to focus fully on landmark detection. By using dilated convolutions to increase the context of the network, the overfitting problems of Pfister *et al*.'s approach are avoided. Note that the successful use of dilated convolutions for global context heavily depends on the implicit kernel convolutions.

3.1. Preprocessing

The training and test data is generated by cropping the images to the (squared) bounding boxes provided by the data sets. The cropped images are then scaled to 96×96px, defensively following the findings of Knoche *et al*. [18], who showed that landmark detection performance typically caps at face resolutions larger than approximately 50×50px. We convert all input images to grayscale since our experiments revealed that performance is on par with or sometimes even superior to RGB or HSV input, probably due to overfitting. This is no requirement for our method, images

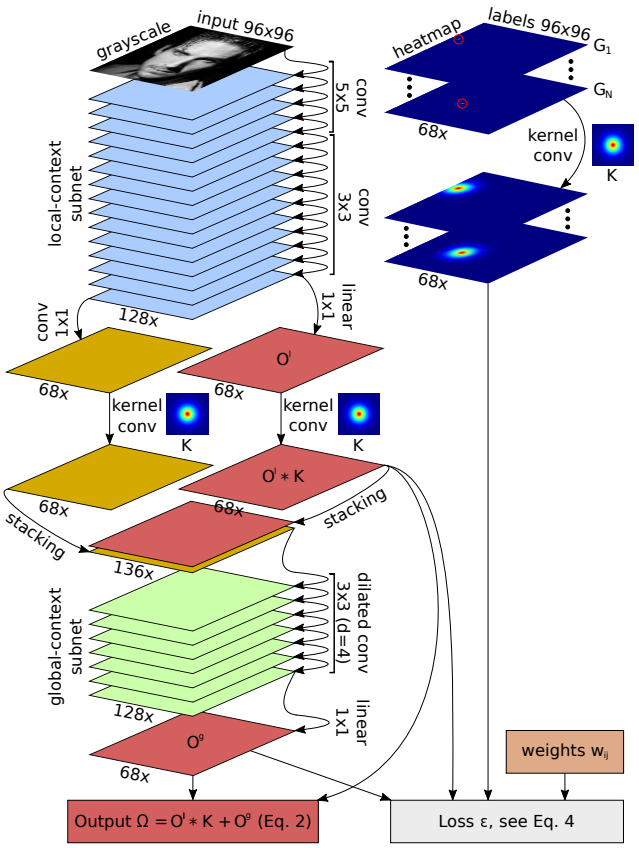

Figure 1. The network architecture used throughout this work.

could be scaled and padded in any fashion. If not for parallelized training in mini-batches, the images could even be of different sizes due to the network being fully convolutional.

Each landmark is represented by a separate heatmap, which can be interpreted as a grayscale image in the range $[0, 1]$. The ground-truth landmark coordinates are set to white, bilinearly interpolated on a black background. In other words, each ground-truth heatmap contains up to 4 non-zero pixels per face that sum up to 1.

3.2. Network architecture

The general structure of our network is illustrated in Figure 1. In summary, the network consists of four consecutive parts, which will be addressed individually in the following subsections:

1. Local-context, fully-convolutional network
2. Convolution with a (customizable) static kernel
3. Global-context, dilated fully-convolutional network
4. Square error-like loss versus kernel-convolved labels

3.2.1 Local-context subnet

The local-context subnet serves as a local detector of facial landmarks from low-level features. On image resolutions such as 96×96px, facial landmarks are already very discriminative on the local level. The local-context subnet thus plays an important role in the overall performance of the network. Our local-context subnet consists of a stack of 15 zero-padded convolutional layers, followed by a linear 1×1 convolution at the end to compensate batch normalization [14].

3.2.2 Kernel convolution

The output of the local-context subnet is convolved with a kernel in channel-wise fashion. The kernel convolution can also be interpreted as a *grouped convolution* [13] with group size 1. It is computed *explicitly* both during training and inference. Nevertheless, we refer to this convolution as *implicit* because it is transparent to the network during backpropagation. So to speak, the local-context subnet does not "know" that a kernel convolution is actually applied, allowing the local-context network to produce clear and sharp outputs.

The kernel convolution serves two main purposes:

1. The pixel-wise square loss now correlates with the *distance* between prediction and ground truth.
2. The global-context subnet can take advantage of dilated rather than dense convolutions.

Without the kernel convolution, the predicted and ground-truth heatmaps are basically discrete indicator functions of the landmark position (neglecting sub-pixel positions). Since the square loss between two indicators is 1 almost everywhere, the slightest mistakes would be penalized just as much as big mistakes. This ultimately leads to the network converging to a state where it is very hesitant to output anything. The kernel acts as a blurring mechanism on those indicator functions. In terms of the loss function to be minimized, the kernel smoothens the surface, reducing local minima by making them more shallow. At the same time, the global minimum's position is unaffected, but being less steep, it is more accessible via gradient descent.

The design of the kernel is flexible and can be used to trade off accuracy versus outlier robustness. We restrict our analysis to the kernel that performed overall best in our experiments, that is, a sum of five Gaussian functions, normalized to a maximum of 1 at the center:

$$K = \frac{2\pi}{5} \sum_{s=1}^{5} (2s-1)^2 \mathcal{N}_2\Big(0, (2s-1)^2\Big), \qquad (1)$$

where $\mathcal{N}_2(\mu, \sigma^2)$ is the symmetric bivariate normal distribution with mean μ and variance σ^2. For computational

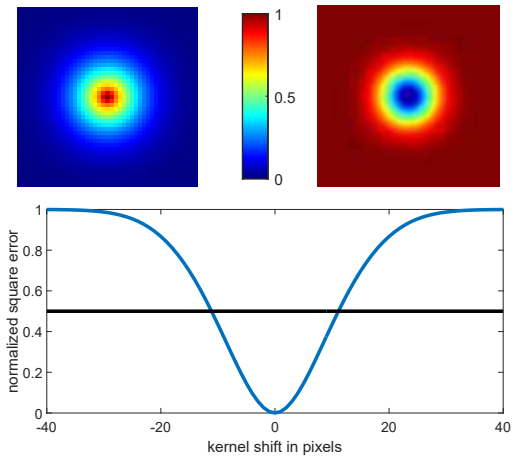

Figure 2. Best viewed in color. Top left: Visualization of the kernel K (Eq. 1). Top right: Square error of two kernels shifted against one another, normalized to twice the kernel square area. Bottom: Centered slice through the top right figure. The horizontal line separates reward (*i.e.*, below threshold) and penalty (*i.e.*, above threshold) for the network.

efficiency, we truncate the kernel to 45×45px. We empirically found that this kernel yields a good trade-off between accuracy and robustness. Figure 2 visualizes the effect of the square error between two such kernels shifted against one another. The network is rewarded for predictions closer or equal to approximately 11px and penalized otherwise. The square error reveals very smooth gradients in this case, which is beneficial for the convergence properties of the network.

3.2.3 Global-context subnet

The global-context subnet's task is to aggregate the output of the local subnet. Pfister *et al*. [24]'s approach is to create a second branch in the network using comparatively large convolutions, for example, 15×15px. Using large convolutions makes the network overfit very easily due to the vast amount of parameters. We avoid overfitting by using dilated convolutions instead. The dilated convolutions increase the receptive field of the network substantially while keeping the number of parameters low.

On the downside, dilated convolutions undersample the input, which creates non-continuous outputs for sparse inputs. In the application of facial landmark detection, a non-continuous output is problematic because there may be a lot of undesirable outlier activations. From a signal processing point of view, dilated convolutions can lead to sampling artifacts, which again may lead to Moiré patterns in the output. Our approach is very resistant against such artifacts because the input is already smoothened by the kernel convolution (*i.e.*, it is low-pass filtered).

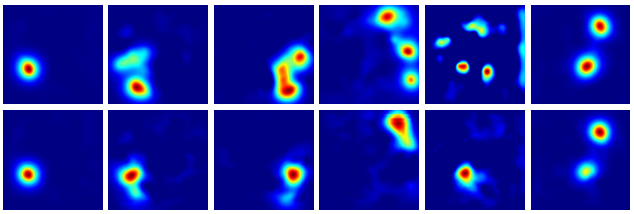

Figure 3. Best viewed in color. Examples of paired heatmap activations of the local-context and full network. Top row: Kernel-convolved output of the local-context subnet, *i.e.*, $K * O^l$. Bottom row: Output Ω of the full network. Note that the output of the global-context subnet can be expressed in terms of both: $O^g = \Omega - K * O^l$ (Eq. 2). In practice, most of the outputs look like the leftmost image.

As mentioned in Section 3.2.1, local features are very discriminative. Consequently, the local-context subnet alone performs relatively well already. Therefore, instead of having the global-context subnet predict the landmark positions from scratch, it is only supposed to *refine* the predictions of the local-context subnet. This is achieved by summation of the local output O^l and the global output O^g in the final network output Ω:

$$\Omega = O^l * K + O^g. \tag{2}$$

The rationale behind our approach is that the global-context subnet should not focus on pixel accuracy but rather on the big picture. For example, the local-context subnet may produce two distinct "right eye" activations; maybe it could not distinguish the left from the right eye, or maybe there are two faces on the image. The dilated convolutions permit the network to check whether there were conflicting (*e.g.*, "left eye") or supportive activations (*e.g.*, "right eyebrow") in the vicinity. The gathered information is then used to reduce or reinforce those activations accordingly.

Figure 3 juxtaposes the kernel-convolved output of the local-context subnet with the final output. While the local-context subnet is able to detect many landmarks on its own, it is confused a number of times and cannot find unique landmark locations. The global-context subnet is able to consider all output channels of the local-context subnet simultaneously, but with a much broader field of view. In situations where the local features are not discriminative enough, the global-context subnet will jump in and try to find a plausible prediction. Otherwise it is mostly inactive.

Our global-context subnet consists of seven consecutive zero-padded dilated convolutions with dilation factor $d = 4$ (*i.e.*, every 4th pixel is sampled) and kernel sizes 3×3, followed by a linear 1×1 convolution at the end to compensate batch normalization [14]. We also considered using deformable convolutions [7] and found that they achieve similar performance, but at a higher computational cost.

3.2.4 Loss

Training data for facial landmark detection is not always complete, which is problematic during training. Most databases are only partially labeled, for example, if landmarks are heavily occluded. We address this problem by excluding non-labeled landmarks from the loss. Furthermore, we weight each of N landmarks depending on whether they lie within the image boundary:

$$w_{ij} = V(i,j) + \frac{L(i,j)}{10} + \frac{L(i,j)}{2N}\sum_{k=1}^{N} V(i,k), \quad (3)$$

where $V(i,j)$ and $L(i,j)$ are indicator functions with $L(i,j) = 1$ iff landmark j in face i is labeled and $V(i,j) = 1$ iff landmark j in face i is within the image boundary, zero otherwise. Note that $L(i,j) = 0$ implies $V(i,j) = 0$. The loss function of the network is then computed as a weighted square error:

$$\epsilon_i = \sum_{n=1}^{N} w_{in}\Big(((O_n^l - G_n) * K)^2 + (O_n^g - G_n * K)^2\Big). \quad (4)$$

O_n^l and O_n^g are the nth channels of the local- and global-context subnet outputs, respectively. K is the kernel (Eq. 1) and G_n the ground truth heatmap for landmark n. For efficiency, the kernel convolution after stacking is reused for the loss. Only O^g is used for inference, but including O^l in the loss function proved to have a positive regularization effect on the network. This is intuitive given that the global-context subnet starts off from whatever the local-context subnet generates.

We also tested a pixel-wise cross entropy loss for the heatmaps, using an additional channel for background such that probabilities sum up to 1. The overall prediction performance on 300-W [28] was worse using cross entropy. However, preliminary experiments with the Kaggle facial keypoint challenge data set [16] suggested that cross entropy outperforms square error by a considerable amount. Kaggle provides comparatively few facial landmarks (*i.e.*, 15); although we did not find conclusive evidence, we reason that cross entropy works better for small numbers of landmarks. On the one hand, cross entropy creates beneficial competition between landmarks. On the other hand, if the landmarks are too crowded, the competition may become too strong, impairing performance.

3.3. From heatmaps to coordinates

In order to compute the inter-pupil normalized error, landmark coordinates must be extracted from the heatmaps created by the network. Since our approach is not restricted to tightly cropped images, there may be multiple faces and/or truncated faces to account for. A simple maximum search is therefore only adequate in benchmark settings where the number of visible landmarks (and faces) is known a priori.

The most trivial approach is to apply a threshold and consider the channel-wise local maxima as positive detections. We found that this works pretty well already, but can sometimes result in outliers. So instead, we interpret the output heatmaps as likelihoods and fit an outlier-robust PCA-based 2D shape model. The model is able to recover from false detections, but also reconstruct occluded or truncated landmarks to a reasonable extent.

For fitting, we first find the approximate rotation of the face using Kabsch's algorithm [15] on the heatmap activations. Next, we generate a linear system of equations, where each heatmap pixel accounts for an equation, weighted by the pixel value. The equations are solved via linear least squares. Only those equations with weights greater than 1% per landmark are considered to save computations.

To reduce the impact of outliers, the fitting result is iteratively refined by considering only those pixels within a certain radius around the prediction. In total, we run three iterations: Global, 20px, and 7px. The model is trained on the same training sets as the network.

4. Experiments

We provide all configurations, models, and code at https://www.mmk.ei.tum.de/cvpr2018/. The network is trained with Microsoft's Cognitive Toolkit (CNTK) [38] in minibatches of 20 images using stochastic gradient descent with momentum. Especially at the beginning of training, the gradients are very large due to the high number of sparsely activated heatmap pixels. We therefore apply gradient clipping with a threshold of 10% in order to avoid exploding gradients. Except for the output layers, all layers are subject to batch normalization [14] with ReLU activations and 10% dropout. On a GTX 1080 Ti, training runs at about 20 samples per second; training 75 epochs on 300-W takes about 2.5 hours per augmentation.

We use 90% of the data for training and 10% for validation. After splitting, the training data is augmented via horizontal mirroring, $\pm 12.5\%$ scaling, and $\pm 30°$ rotation, increasing the training set size by a factor of 18. The model that performs best on the validation set is then selected for testing. Specifically for Menpo [41], since no bounding boxes are provided, we take the tight (squared) bounding box around the ground truth landmarks, enlarged by 17%.

For a fair comparison with other works, we deliberately refrain from using computationally expensive post-processing methods such as model-averaging via test-time dropout [9] or test-time data augmentation. We found that the error can be reduced by $5 - 10\%$ (relative) on average with a tenfold ensemble using either of those methods.

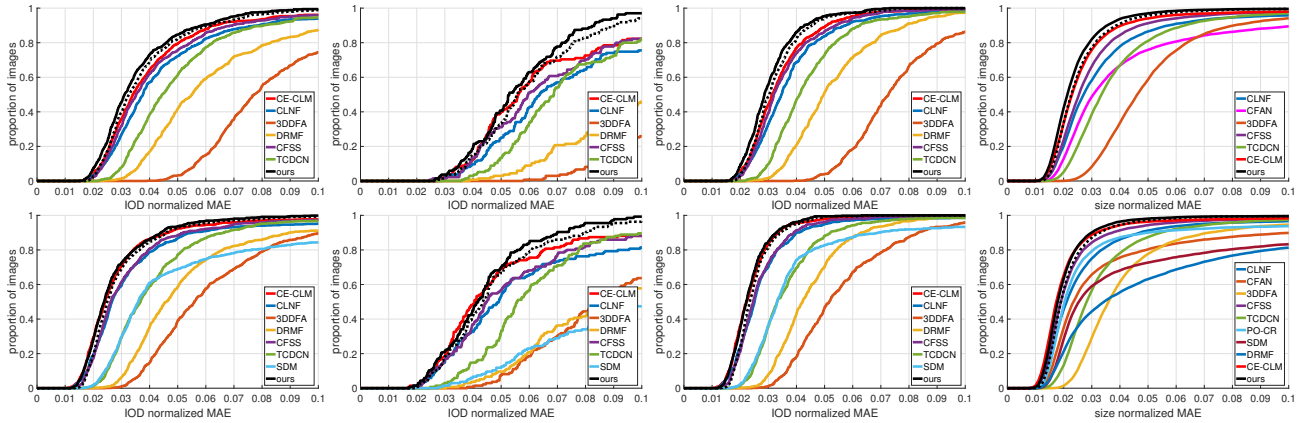

Figure 4. Best viewed in color. Quantitative results of our approach trained on the 300-W training set. Top/bottom row: With/without face outline. From left to right: 300-W [28], iBUG [29], LFPW [5] + HELEN [20], Menpo frontal train set [41]. The dashed line depicts the performance of our network without model fitting (*i.e.*, simple heatmap-wise maximum). Our approach is clearly more robust against outliers than other methods, most prominently on the challenging iBUG test set. Most of the time we are also more accurate.

4.1. Quantitative results

We evaluate our approach using the code provided by Zadeh *et al*. [40]. Where possible, we compare against the model-based approaches CLNF [4] and 3DDFA [45] (Section 2.1) as well as the regression-based approaches TCDCN [43], CFSS [44], SDM [37], CFAN [42], PO-CR [35], DRMF [1], and CE-CLM [40] (Section 2.2). Figure 4 illustrates the results for inter-ocular distance (IOD) normalized mean absolute error (MAE) on the 300-W benchmark [28] and on a cross-data set test with Menpo [41]. Note that we use fewer data augmentations during training than CE-CLM ($\times 18$ vs. $\times 56$). Furthermore, some works train on slightly different data than 300-W, for example, Multi-PIE [10] instead of AFW [46] (*e.g.*, CE-CLM [40]). Nevertheless, the differences are relatively marginal so that the comparison is still fair overall.

From the top row of Figure 4, most notably, our approach is much more outlier robust than other methods, especially for the very challenging iBUG test set. Only 4 of the 689 300-W test images have an error of more than 10% IOD, whereas the next-most robust algorithm CE-CLM [40] fails for 28 images. At the same time, we outperform the other methods over most of the spectrum. Interestingly, not much performance is gained from the 2D PCA-based shape model, which is proof that our local-global context network indeed learned to interpret global context on its own.

The face outline is considered to be very challenging due to the high variation and inaccurate ground truth labels. When the face outline is not considered (Figure 4, bottom row), we are still better, but the margin to CE-CLM shrinks. We derive that our model deals very well with label variance and succeeds in finding good face outlines even under occlusion.

Following Zadeh *et al*. [40], we also report the median landmark error on 300-W and the Menpo cross-data set test in Table 1. Comparing the median error, we still perform overall better than CE-CLM by Zadeh *et al*. [40], but the differences are less significant compared to Figure 4. This indicates that the local patch experts of Zadeh and colleagues provide better local predictions than our local-context subnet.

4.2. Qualitative results

Some qualitative results on the 300-W challenge data set [28] are presented in Figure 5. The top four rows show a selection of images with difficult poses, expressions, lighting conditions, and/or occlusions. The results demonstrate that our approach is very robust even under difficult conditions. The bottom row depicts the worst 10 results from the benchmark set. Arguably, not all among the 10 worst results are failure cases, which is partly due to debatable ground truth labels, especially for the face outlines. The most challenging images are profile images with self-occlusions. This is mostly because the PCA-based 2D model is not able to accurately represent 3D rotational information and distorts the shape of the face near the occlusion, for example, in the 5th image of the bottom row.

4.3. Face detection-less landmark detection

While the benchmark results so far revealed superior performance compared to other methods, the experiments were done in a controlled setting: Only one face per image with uniform image and face resolutions. In the wild, this is usually achieved via a robust face detector. One benefit of our approach compared to other methods is that no face detector is required. In this section, we present a proof of concept for this claim.

Method \ Data	iBUG [29]	LFPW [5] + HELEN [20]	Menpo [41] (frontal)
CLNF [4]	6.37/4.93	3.47/2.51	2.66/2.10
SDM [37]	– /10.73	– /3.31	– /2.54
CFAN [42]	8.38/6.99	– /–	2.87/2.34
DRMF [1]	10.36/8.64	4.97/4.22	– /3.44
CFSS [44]	5.97/4.49	3.20/2.46	2.32/1.90
TCDCN [43]	6.87/5.56	4.11/3.32	3.32/2.81
3DDFA [45]	12.31/8.34	7.27/5.17	4.51/3.59
PO-CR [35]	– /**3.33**	– /2.67	– /2.03
CE-CLM [40]	5.62/4.05	3.13/2.23	2.23/**1.74**
Ours (no model fit)	5.55/4.36	3.04/2.34	2.27/1.90
Ours	**5.29**/4.18	**2.86**/**2.21**	**2.14**/1.79

Table 1. Median IOD-normalized MAE with/without face outline for iBUG [29] and LFPW [5] + HELEN [20]. Median image size-normalized MAE with/without face outline for Menpo [41]. The best performance is highlighted in bold. While our approach does not achieve the best median performance on all datasets, the performance is very consistent.

Figure 5. Best viewed in the digital version. Qualitative results of our approach on the 300-W benchmark. The images are sorted according to their error (top left is best, bottom right is worst). All but the first two images shown are worse than the average case. More precisely, the mean and median errors of the images in rows 1 through 4 are in the 76.5% and 83.6% quantiles of the test set, respectively. The 5th row displays the 10 worst results. Note that the results are displayed in color, but our network only uses grayscale information.

First, the network is trained with one slight modification: We add 30 000 non-face images from IJB-B [17] to the training set. This is not a strict requirement, but without the non-face background images from IJB-B, there would be more outlier activations in the background. Everything else remains unchanged. Weighting the loss according to Equation 3 is key when training with background images, otherwise the network focuses too much on background rather than faces.

With this adapted network in place, we manually select images with multiple faces at multiple scales from AFLW [19] and feed them through the network. We start at the original resolution, iteratively reducing resolution by factors of two. The process stops with the first resolution below 96×96px. For example, a 640×480px image would require four iterations, the last one being 80×60px. The heatmap outputs are then scaled up to the original resolution and combined via a pixel-wise hard maximum. Local

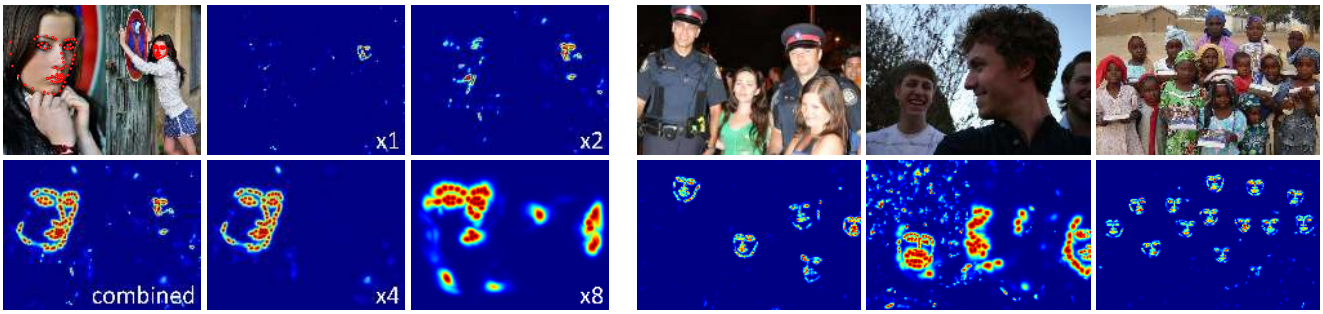

Figure 6. Best viewed in the digital version. Proof of concept for our face detection-less landmark detection on a selection of AFLW images [19]. Left: Full pipeline with resulting landmark predictions. Right: Three further examples. For the purpose of better visualization, we normalize the heatmaps to a maximum of 1 per channel and enhance the contrast. We compute and display the per-pixel maximum of the heatmaps generated by our local-global context network at different scales of the input image ($\times 1 - \times 8$). The *combined* heatmap is again the pixel-wise maximum of all four resampled outcomes. After clustering the groups of candidate detections into regions of interest, we apply a 2D PCA-based shape model to the non-normalized heatmaps, yielding the final landmark predictions.

maxima above a simple threshold within each heatmap represent candidate detections. The candidate detections are clustered to generate face hypotheses which are verified or rejected by fitting the PCA-based shape model from Section 3.3. Some qualitative results using this approach are illustrated in Figure 6.

5. Conclusion

We proposed a new network architecture for facial landmark detection via fully-convolutional heatmap regression. The core concept is an implicit kernel convolution between a local-context subnet and a global-context subnet composed of dilated convolutions. The local-context subnet is responsible for landmark proposals, which are refined by the global-context subnet. In a postprocessing step, a PCA-based 2D shape model is fitted to the heatmaps generated by the network in order to retrieve landmark coordinates. Our approach beats the state of the art on 300-W [28] and on a cross-data set test with Menpo [41]. We demonstrated that our network, in contrast to other methods, does not require a face detector and can handle multiple faces at multiple resolutions.

Although we apply our approach specifically to facial landmark detection, the concept can be generalized to any heatmap-like regression tasks, for example, foreground-background segmentation [2], object segmentation [39], human pose estimation [6, 24, 34], etc. Specifically, networks building on fully-convolutional architectures such as U-Net [27] or SegNet [3] may profit from our approach.

References

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3444–3451, 2013. 2, 6, 7

[2] M. Babaee, D. T. Dinh, and G. Rigoll. A deep convolutional neural network for background subtraction. *arXiv preprint arXiv:1702.01731*, 2017. 2, 8

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015. 2, 8

[4] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Continuous conditional neural fields for structured regression. In *European Conference on Computer Vision*, pages 593–608. Springer, 2014. 2, 6, 7

[5] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2930–2940, 2013. 6, 7

[6] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*, pages 717–732. Springer, 2016. 2, 8

[7] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. *arXiv preprint arXiv:1703.06211*, 2017. 4

[8] H. Fan and E. Zhou. Approaching human level facial landmark localization by deep learning. *Image and Vision Computing*, 47:27–35, 2016. 1

[9] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016. 5

[10] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 6

[11] R. A. Güler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou, and I. Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6799–6808, 2017. 2

[12] Z. He, J. Zhang, M. Kan, S. Shan, and X. Chen. Robust fec-cnn: A high accuracy facial landmark detection system. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, volume 3, page 6, 2017. 2

[13] Y. Ioannou, D. Robertson, R. Cipolla, and A. Criminisi. Deep roots: Improving cnn efficiency with hierarchical filter groups. *arXiv preprint arXiv:1605.06489*, 2016. 3

[14] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015. 3, 4, 5

[15] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976. 5

[16] Kaggle. Facial keypoint detection competition. 2016. 5

[17] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1931–1939, 2015. 7

[18] M. Knoche, D. Merget, and G. Rigoll. Improving facial landmark detection via a super-resolution inception network. In *German Conference on Pattern Recognition*, pages 239–251. Springer, 2017. 2

[19] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *IEEE International Conference on Computer Vision Workshops*, pages 2144–2151, 2011. 7, 8

[20] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *European Conference on Computer Vision*, pages 679–692. Springer, 2012. 6, 7

[21] W. Luo, Y. Li, R. Urtasun, and R. Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 4898–4906, 2016. 1

[22] B. Martinez and M. F. Valstar. Advances, challenges, and opportunities in automatic facial expression recognition. In *Advances in Face Detection and Facial Image Analysis*, pages 63–100. Springer, 2016. 1

[23] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015. 2

[24] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1913–1921, 2015. 1, 2, 4, 8

[25] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[26] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 1

[27] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 8

[28] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18, 2016. 2, 5, 6, 8

[29] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013. 6, 7

[30] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011. 2

[31] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013. 2

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 1

[33] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014. 1

[34] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems*, pages 1799–1807, 2014. 2, 8

[35] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3659–3667, 2015. 2, 6, 7

[36] S. Xiao, J. Li, Y. Chen, Z. Wang, J. Feng, S. Yan, and A. Kassim. 3D-assisted coarse-to-fine extreme-pose facial landmark detection. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, volume 3, page 6, 2017. 2

[37] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539, 2013. 2, 6, 7

[38] D. Yu, A. Eversole, M. Seltzer, K. Yao, Z. Huang, B. Guenter, O. Kuchaiev, Y. Zhang, F. Seide, H. Wang, et al. An introduction to computational networks and the computational network toolkit. *Microsoft Technical Report MSR-TR-2014–112*, 2014. 5

[39] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 2, 8

[40] A. Zadeh, Y. C. Lim, T. Baltrušaitis, and L.-P. Morency. Convolutional experts constrained local model for 3D facial

landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2519–2528, 2017. 1, 2, 6, 7

[41] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *Computer Vision and Pattern Recognition Workshop*, 2017. 2, 5, 6, 7, 8

[42] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European Conference on Computer Vision*, pages 1–16. Springer, 2014. 2, 6, 7

[43] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, 2014. 2, 6, 7

[44] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015. 2, 6, 7

[45] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3D solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 146–155, 2016. 2, 6, 7

[46] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, 2012. 6