DenseReg: Fully Convolutional Dense Shape Regression In-the-Wild

Rıza Alp Güler [1] George Trigeorgis [2] Epameinondas Antonakos [2]
Patrick Snape [2] Stefanos Zafeiriou [2] Iasonas Kokkinos [3]
[1]INRIA-CentraleSupélec, France [2]Imperial College London, UK [3]University College London, UK
[1]riza.guler@inria.fr [2]{g.trigeorgis, e.antonakos, p.snape,s.zafeiriou}@imperial.ac.uk [3]i.kokkinos@cs.ucl.ac.uk

Abstract

In this paper we propose to learn a mapping from image pixels into a dense template grid through a fully convolutional network. We formulate this task as a regression problem and train our network by leveraging upon manually annotated facial landmarks "in-the-wild". We use such landmarks to establish a dense correspondence field between a three-dimensional object template and the input image, which then serves as the ground-truth for training our regression system. We show that we can combine ideas from semantic segmentation with regression networks, yielding a highly-accurate 'quantized regression' architecture.

Our system, called DenseReg, allows us to estimate dense image-to-template correspondences in a fully convolutional manner. As such our network can provide useful correspondence information as a stand-alone system, while when used as an initialization for Statistical Deformable Models we obtain landmark localization results that largely outperform the current state-of-the-art on the challenging 300W benchmark. We thoroughly evaluate our method on a host of facial analysis tasks, and also demonstrate its use for other correspondence estimation tasks, such as modelling of the human ear. DenseReg code is made available at http://alpguler.com/DenseReg.html along with supplementary materials.

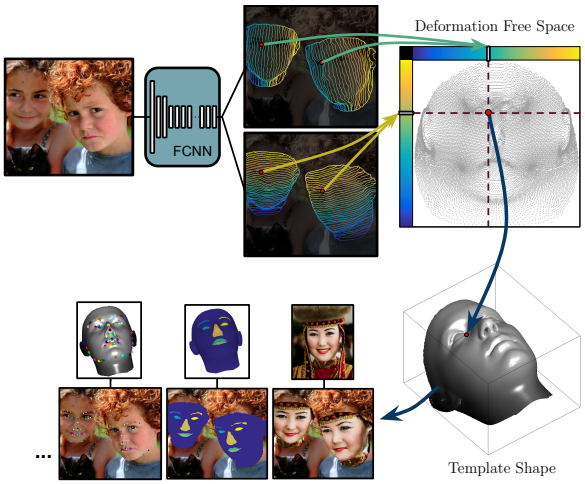


Figure 1: We introduce a fully convolutional neural network that regresses from the image to a "canonical", deformation-free parameterization of the face surface, effectively yielding a dense 2D-to-3D surface correspondence field. Once this correspondence field is available, one can effortlessly solve many image-level problems by backward-warping their canonical solution from the template coordinates to the image domain for the problems of landmark localization, semantic part segmentation, and face transfer.

1. Introduction

Non-planar object deformations, e.g. due to facial pose or expression, result in challenging but also informative signal variations. Our objective in this paper is to recover this information in a feedforward manner by employing a discriminatively trained convolutional network. Our motivation for this is the understanding that there is a gap between discriminatively trained systems for detection and category-level deformable models; we propose a system that combines the merits of both.

In particular, discriminative learning-based approaches typically pursue invariance to shape deformations, for instance by employing local 'max-pooling' operations to elicit responses that are invariant to *local* translations. As such, these models can reliably detect patterns irrespective of their deformations through efficient, feedforward algorithms. At the same time however this discards useful shape-related information and only delivers a single categorical decision per position. Several recent works in deep learning have aimed at enriching deep networks with information about shape by explicitly modelling *the effect* of similarity transformations [31] or non-rigid deformations [20, 18, 9]; several of these have found success in classification [31], fine-grained recognition [20], and also face detection [9]. There are works [24, 33] that model the de-

formation via optimization procedures, whereas we obtain it in a feedforward manner and in a single shot. In these works, shape is treated as a nuisance, while we treat it as the goal in itself. Recent works on 3D surface correspondence [29, 5] have shown the merit of CNN-based unary terms for correspondence. In our case we tackle the much more challenging task of establishing a 2D to 3D correspondence *in the wild* by leveraging upon recent advances in semantic segmentation [10]. To the best of our knowledge, the task of explicitly recovering dense correspondence in the wild has not been addressed yet in the context of deep learning.

By contrast, approaches that rely on Statistical Deformabe Models (SDMs), such as Active Appearance Models (AAMs) or 3D Morphable Models (3DMMs) aim at explicitly recovering dense correspondences between a deformation-free template and the observed image, rather than trying to discard them. This allows one to both represent shape-related information (*e.g.* for facial expression analysis) and also to obtain invariant decisions after registration (*e.g.* for identification). Explicitly representing shape can have substantial performance benefits, as is witnessed in the majority of facial analysis tasks requiring detailed face information e.g. landmark localisation [35], 3D pose estimation, as well as 3D face reconstruction "in-the-wild" [22], where SDMs consitute the current state of the art.

However SDM-based methods are limited in two respects. Firstly they require an initialization from external systems, which can become increasingly challenging for elaborate SDMs: AAMs may require only a bounding box type of initialisation but 3DMMs further require a good initialization of the position of particular facial landmarks - while SDM performance may drop due to poor initialization. Furthermore, SDM fitting requires iterative, time-demanding optimization algorithms, especially when the initialisation is far from the solution. The advent of Deep Learning has made it possible to replace the iterative optimization task with iterative regression problems [38], but this does not alleviate the need for initialization and multiple iterations.

In this work we aim at bridging these two approaches, and introduce a discriminatively trained network to obtain, in a fully-convolutional manner, dense correspondences between an input image and a deformation-free template coordinate system.

In particular, we exploit the availability of manual facial landmark annotations "in-the-wild" in order to fit a 3D template; this provides us with a dense correspondence field, from the image domain to the 2-dimensional, $U - V$ parameterization of the face surface. We then train a fully convolutional network that densely regresses from the image pixels to this $U - V$ coordinate space.

This provides us with dense and fine-grained correspondence information, as in the case of SDMs, while at the same time being independent of any initialization procedure, as in the case of discriminatively trained 'fully-convolutional' networks. We demonstrate that the performance of certain tasks, such as facial landmark localisation or segmantic part segmentation, is largely improved by using the proposed network.

Even though the methodology is general this paper is mainly concerned with human faces. The general architecture for the case of human face is described in Fig. 1.

Our approach can be on the one hand understood as providing a stand-alone, feedforward alternative to the combination of initialization with iterative fitting typically used in SDMs. This allows us to have a feedforward system that solves both the detection and correspondence problems at approximate $7 - 8$ frames per second for a 300×300 input image. On the other hand, our approach can also be understood as an initialization procedure for SDMs which gets them started from a much more accurate position than the bounding box, or landmark-based initializations currently employed in the face analysis literature. When taking this approach we observe substantial gains over the current state-of-the-art systems.

We can summarize our contributions as follows:

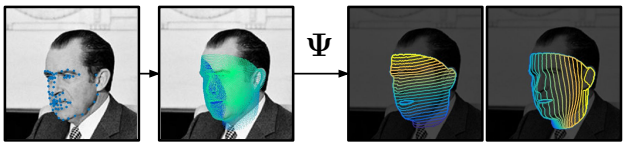
- We introduce the task of dense shape regression in the setting of CNNs, and exploit the SDM-based notion of a deformation-free UV-space to construct target ground-truth signals (Sec.2).

- We propose a carefully-designed fully-convolutional shape regression system that exploits ideas from semantic segmentation and dense regression networks. Our *quantized regression* architecture(Sec.3) is shown to substantially outperform simpler baselines that consider the task as a plain regression problem .

- We use dense shape regression to jointly tackle a multitude of problems, such as landmark localization or semantic segmentation. In particular, the template coordinates allow us to 'copy' multiple annotations constructed on a single template system, and thereby tackle multiple problems in a single go.

- We use the regressed shape coordinates for the initialization of SDMs; systematic evaluations on facial analysis benchmarks show that this yields substantial performance improvements on tasks ranging from landmark localization to semantic segmentation.

- We demonstrate the generic nature of the method by applying it to the task of estimating dense correspondence information for human ears.



(a) (b) (c)

Figure 2: Ground-truth generation: *(a)* Annotated landmarks. *(b)* Template shape morphed based on the landmarks. *(c)* Deformation-free coordinates (u^h and u^v), obtained by unwrapping the template shape, transferred to image domain.

Our system is particularly simple to implement, as it relies on a variation of the broadly adopted Deeplab system [10].

2. From SDMs to Dense Shape Regression

Following the deformable template paradigm [46, 17], we consider that object instances are obtained by deforming a prototypical object, or 'template', through dense deformation fields. This makes it possible to factor object variability within a category into variations that are associated to deformations, generally linked to the object's 2D/3D shape, and variations that are associated to appearance (or, 'texture' in graphics), e.g. due to facial hair, skin color, or illumination.

This factorization largely simplifies the modelling task. SDMs use it as a stepping stone for the construction of parametric models of deformation and appearance. For instance in AAMs a combination of Procrustes Analysis, Thin-Plate Spline warping and PCA is the standard pipeline for learning a low-dimensional linear subspace that captures category-specific shape variability [13]. Even though we have a common starting point, rather than trying to construct a linear generative model of deformations, we treat the image-to-template correspondence as a vector field that our network tries to regress.

In particular, we start from a template $\boldsymbol{X} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, ...\boldsymbol{x}_m^\top]^\top \in \mathbb{R}$, where each $\boldsymbol{x}_j \in \mathbb{R}^3$ is a vertex location of the mesh in 3D space. This template is obtained by the automatic pipeline proposed from Booth et al. [3] which brings a large set of 3D facial scans in correspondence through the use of Procrustes alignment and an adapted Non-Rigid ICP optimisation problem. We compute a bijective mapping ψ, from template mesh $\boldsymbol{X}$ to the 2D canonical space $\boldsymbol{U} \in \mathbb{R}^{2\times m}$, such that

$$\psi(\boldsymbol{x}_j) \mapsto \boldsymbol{u}_j \in \boldsymbol{U} \quad , \quad \psi^{-1}(\boldsymbol{u}_j) \mapsto \boldsymbol{x}_j. \tag{1}$$

The mapping ψ is obtained via the cylindrical unwrapping described in [4]. Thanks to the cylindrical unwrapping, we can interpret these coordinates as being the horizontal and vertical coordinates while moving on the face surface: $u_j^h \in [0,1]$ and $u_j^v \in [0,1]$. Note that this semantically meaningful parameterization has no effect on the operation of our method.

We exploit the availability of landmark annotations "in the wild", to fit the template face to the image by obtaining a coordinate transformation for each vertex $\boldsymbol{x}_j$. This involves estimating the morphable model parameters and weak perspective projection parameters following [22]. The corresponding canonical coordinate $\boldsymbol{u}_j$ for each vertex on the template face is then transferred to the morphed 3D shape. The canonical coordinates that correspond to the visible image pixels are then obtained in 2D by a z-buffering operation. As illustrated in Fig. 2, once the transformation from the template face vertices to the morphed vertices is established, the $\boldsymbol{u}_j$ coordinates of each visible vertex on the canonical face can be transferred to the image space. This establishes the ground truth signal for our subsequent regression task. We intend to make the established correspondences publicly available.

3. Fully Convolutional Dense Shape Regression

Having described how we establish our supervision signal, we now turn to the task of estimating it through a convolutional neural network (CNN). Our aim is to estimate at any image pixel that belongs to a face region the values of $\boldsymbol{u} = [u^h, u^v]$. We need to also identify non-face pixels, e.g. by predicting a 'dummy' output.

On the one hand one can phrase this problem as a generic regression task and attack it with the powerful machinery of CNNs. Unfortunately, the best performance that we could obtain this way was quite underwhelming, apparently due to the task's complexity. Our approach is to quantize and estimate the quantization error separately for each quantized value. Instead of directly regressing u, the quantized regression approach lets us solve a set of easier sub-problems, yielding improved regression results.

In particular, instead of using a CNN as a 'black box' regressor, we draw inspiration from the success of recent works on semantic part segmentation [39, 11], and landmark classification [6, 7]. These works have shown that CNNs can deliver remarkably accurate predictions when trained to predict *categorical variables*, indicating for instance the facial part or landmark corresponding to each pixel.

Building on these successes, we propose a hybrid method that combines a classification with a regression problem. Intuitively, we first identify a coarser face region that can contain each pixel, and then obtain a refined, region-specific prediction of the pixel's $U-V$ field. As we will describe below, this yields substantial gains in performance when compared to the baseline of a generic regression system.

We identify facial regions by using a simple geometric approach. We tesselate the template's surface with a carte-

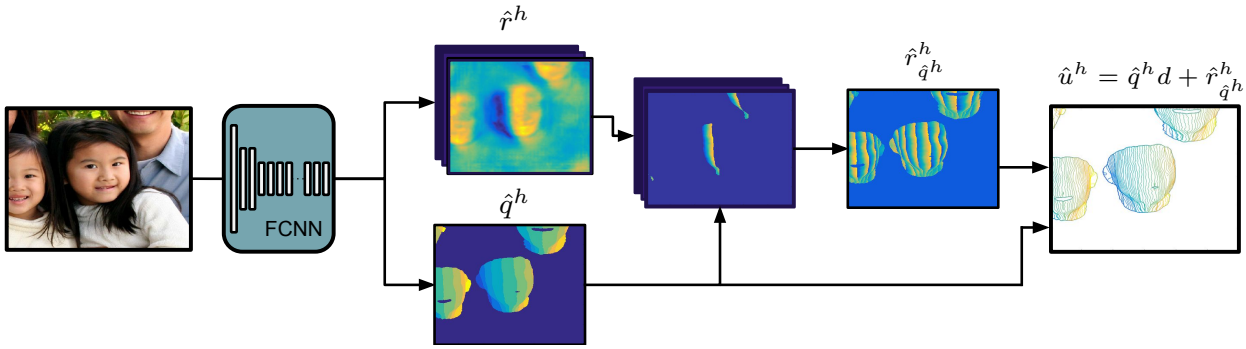


Figure 3: Proposed Quantized Regression Approach for the horizontal correspondence signal: The continuous signal is regressed by first estimating a grossly quantized (or, discretized) function through a classification branch. For each quantized value $\hat{q}^h$ we use a separate residual regression unit's prediction, $\hat{r}^h_{\hat{q}^h}$, effectively multiplexing the different residual predictions. These are added to the quantized prediction, yielding a smooth and accurate correspondence field.

sian grid, by uniformly and separately quantizing the u^h and u^v coordinates into K bins, where K is a design parameter. For any image that is brought into correspondence with the template domain, this induces a discrete labelling, which can be recovered by training a CNN for classification.

On Fig. 4, the tesselations of different granularities are visualized. For a sufficiently large value of K even a plain classification result could provide a reasonable estimate of the pixel's correspondence field, albeit with some staircasing effects. The challenge here is that as the granularity of these discrete labels becomes increasingly large, the amount of available training data decreases and label complexity increases. A more detailed analysis on the effect of label-space granularity to segmentation performance is provided in supplementary materials.

We propose to combine powerful classification results with a regression problem that will yield a refined correspondence estimate. For this, we compute the residual between the desired and quantized $U-V$ coordinates and add a separate module that tries to regress it. We train a separate regressor per facial region, and at any pixel only penalize the regressor loss for the responsible face region. We can interpret this form as a 'hard' version of a mixture of regression experts [21]. This interpretation is further elaborated upon in the supplementary material.

The horizontal and vertical components u^h, u^v of the correspondence field are predicted separately. This results

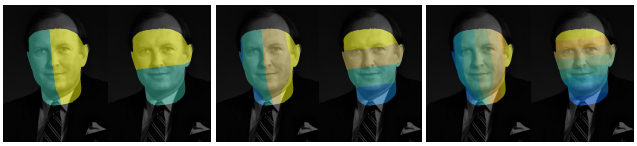


Figure 4: Horizontal and vertical tesselations obtained using $K = 2, 4$ and 8 bins.

in a substantial reduction in computational and sample complexity - For K distinct U and V bins we have K^2 regions; the classification is obtained by combining 2 K-way classifiers. Similarily, the regression mapping involves K^2 regions, but only uses $2K$ one-dimensional regression units. The pipeline for quantized face shape regression is provided in Fig. 3.

We now detail the training and testing of this network; for simplicity we only describe the horizontal component of the mapping. From the ground truth construction, every position $\boldsymbol{x}$ is associated with a scalar ground-truth value u^h. Rather than trying to predict u^h as is, we transform it into a pair of discrete q^h and continuous r^h values, encoding the quantization and residual respectively:

$$q^h = \lfloor \frac{u^h}{d} \rfloor, \quad r^h_i = \left(u^h_i - q^h_i d\right), \tag{2}$$

where $d = \frac{1}{K}$ is the quantization step size (we consider u^h, u^v coordinates to lie in $[0, 1]$).

Given a common CNN trunk, we use two classification branches to predict q^h, q^v and two regression branches to predict r^h, r^v as convolution layers with kernel size 1×1. As mentioned earlier, we employ separate regression functions per region, which means that at any position we have K estimates of the horizontal residual vector, $\hat{r}^h_i,\ i = 1, \ldots, K$.

At test time, we let the network predict the discrete bin $\hat{q}^h$ associated with every input position, and then use the respective regressor output $\hat{r}^h_{\hat{q}^h}$ to obtain an estimate of u:

$$\hat{u}^h = \hat{q}^h d + \hat{r}^h_{\hat{q}^h} \tag{3}$$

For the q^h and q^v, which are modeled as categorical distributions, we use softmax followed by trcross entropy loss. For estimating $\hat{r}^h$ and $\hat{r}^v$, we use a normalized version of the smooth L_1 loss in [16]. The normalization is obtained

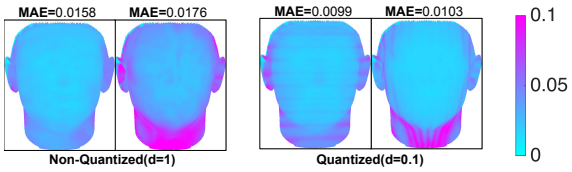


Figure 5: Visualization of local average absolute error on pairs of vertical-horizontal coordinate regression pairs for both quantized and non-quantized regression. We visualize the mean absolute error for each approach. The proposed quantized regression clearly outperforms plain regression.

by dividing the loss by the number of pixels that contribute to the loss.

Compared to plain regression of the coordinates, the proposed method achieves much better results. For a clear demonstartion of this, we resort to two main branches of analysis. Firstly, we decouple the effect of foreground/background segmentation, by only analysing the error of detected foreground pixels. We transfer absolute errors (absolute value of the difference between estimated and regressed coordinates) to the template model using the ground-truth coordinates. The errors corresponding to each vertex are visualized in Fig. 5 along with mean absolute error computed from all of the detected pixels. We can observe that the errors are concentrated at the neck region, where the fits might be rather inconsistent due to the lack of landmarks. We can also observe that the quantized regression approach yields much smaller errors, concentrated mainly at boundaries of quantized regions. Secondly, we plot the Cummulative Error Distribution(CED) for both of the approaches on Fig. 6. The error is normalized by the distance between two eyes in the deformation-free coordinate system, to analyze the error in the range that is commonly used in facial landmark analysis. The results clearly show that the quantized approach is significantly better, especially at the 0 - 0.1 regime, which is relevant for applications.

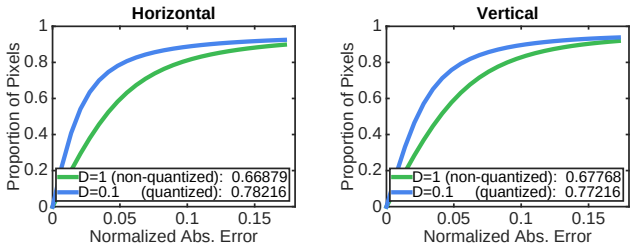


Figure 6: Cumulative Error Distribution of absolute errors normalized by the interocular distance on the deformation-free coordinate system for quantized and non-quantized regression approaches.The reported Area Under the Curve (AUC) indicates that our alternative to the non-quantized regression task performs substantially better.

4. Experiments

Herein, we evaluate the performance of the proposed method (referred to as `DenseReg`) on various face-related tasks. In the following sections, we first describe the training setup (Sec. 4.1) and then present extensive quantitative and qualitative results on *(i)* semantic segmentation (Sec. 4.2), *(ii)* landmark localization on static images (Sec. 4.3), *(iii)* deformable tracking (Sec. 4.4) , *(iv)* monocular depth estimation (Sec. 4.5) and *(v)* human ear landmark localization (Sec. 4.6). We note that DenseReg is not trained or fine-tuned specifically for each one of those tasks. Its predictions are used out-of-the-box for different task-specific evaluations.

4.1. Training Setup

Training Databases. We train our system using the 300W database [36, 35] that is annotated with 68 landmark points. Our training set consists of the LFPW trainset [2], Helen trainset [25] and AFW [49], thus 3148 images that are captured under completely unconstrained conditions and exhibit large variations in pose, expression, illumination, age, etc. Many of these images contain multiple faces, some of which are not annotated. We deal with issue by employing the out-of-the-box DPM face detector of Mathias et al. [30] to obtain the regions that contain a face for all of the images. The detected regions that do not overlap with the ground truth landmarks do not contribute to the loss. For training and testing, we have rescaled the images such that their largest side is 800 pixels.

CNN Training. For the dense regression network, we adopt a ResNet101 [19] architecture with dilated convolutions (atrous) [10, 27], such that the stride of the CNN is 8. We use bilinear interpolation to upscale both the $\hat{q}$ and $\hat{r}$ branches before the losses. The losses are applied at the input image scale and back-propagated through interpolation. We apply a weight to the smooth $L1$ loss layers to balance their contribution. In our experiments, we have used a weight of 40 for quantized ($d = 0.1$) and a weight of 70 for non-quantized regression, which are determined by a coarse cross validation. We initialize the training with a network pre-trained for the MS COCO segmentation task [26]. The new layers are initialized with random weights drawn from Gaussian distributions. Large weights of the regression losses can be problematic at initialization even with moderate learning rates. To cope with this, we use initial training with a lower learning rate for a *warm start*. We then use a base learning rate of 0.001 with a polynomial decay policy for $20k$ iterations with a batch size of 10 images. During training, each sample is randomly scaled with one of the ratios $[0.5, 0.75, 1, 1.25, 1.5]$ and cropped to form a fixed 321×321 input image.

Class	*Methods*	
	DenseReg	Deeplab-v2
Left Eyebrow	48.35	40.57
Right Eyebrow	46.89	41.85
Left Eye	75.06	73.65
Right Eye	73.53	73.67
Upper Lip	69.52	62.04
Lower Lip	75.18	70.71
Nose	87.71	86.76
Other	99.44	99.37
Average	**71.96**	68.58

Table 1: Semantic segmentation accuracy on Helen testset measured using intersection-over-union (IoU) ratio.

4.2. Semantic Segmentation

As discussed in Sec. 2, any labelling function defined on the template shape can be transferred to the image domain using the regressed coordinates. One application that can be naturally represented on the template shape is semantic segmentation of facial parts. To this end, we manually defined a segmentation mask of 8 classes (right/left eye, right/left eyebrow, upper/lower lip, nose, other) on the template shape, as shown in Fig. 1.

We compare against a state-of-the-art semantic part segmentation system (DeepLab-v2) [11] which is based on the same ResNet-101 architecture as our proposed DenseReg. We train DeepLab-v2 on the same training images (i.e. LFPW trainset, Helen trainset and AFW). We generate the ground-truth segmentation labels for both training and testing images by transferring the segmentation mask using the ground-truth deformation-free coordinates explained in Sec. 2. We employ the Helen testset [25] for the evaluation.

Table 1 reports evaluation results using the intersection-over-union (IoU) ratio. Additionally, Fig. 13 shows some qualitative results for both methods, along with the ground-truth segmentation labels. The results indicate that the DenseReg outperforms DeepLab-v2. The reported improvement is substantial for several parts, such as eyebrows and lips. We believe that this result is significant given that DenseReg is not optimized for the specific task-at-hand, as opposed to DeepLab-v2 which was trained for semantic segmentation. This performance difference can be justified by the fact that DenseReg was exposed to a richer label structure during training, which reflects the underlying variability and structure of the problem.

4.3. Landmark Localization on Static Images

DenseReg can be readily used for the task of facial landmark localization on static images. Given the landmarks' locations on the template shape, it is straightforward to estimate the closest points in the deformation-free coordinates on the images. The local minima of the Euclidean distance

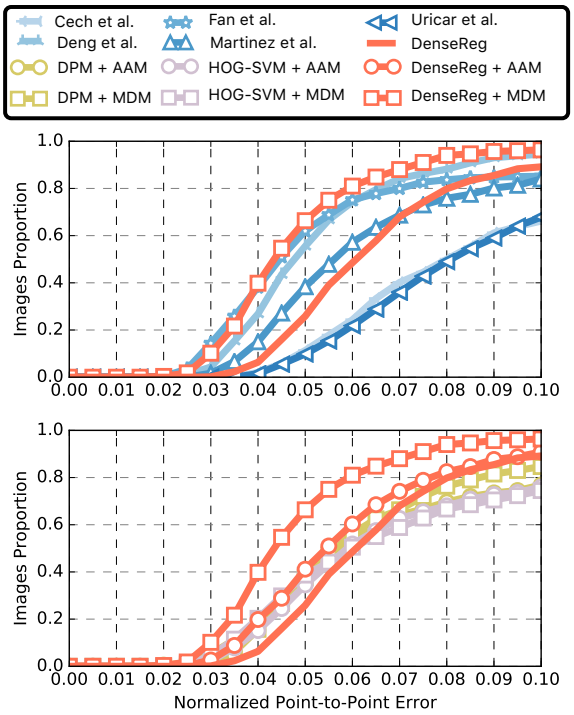


Figure 7: Landmark localization results on the 300W testing dataset using 68 points. Accuracy is reported as Cumulative Error Distribution of RMS point-to-point error normalized with interocular distance. *Top:* Comparison with state-of-the-art. *Bottom:* Self-evaluation results.

between the estimated coordinates and the landmark coordinates are considered as detected landmarks. In order to find the local minima, we simply analyze the connected components separately. Even though more sophisticated methods for covering "touching shapes" can be used, we found that this simplistic approach is sufficient for the task.

Note that the closest deformation-free coordinates among all *visible* pixels to a landmark point is not necessarily the correct corresponding landmark. This phenomenon is called "landmark marching" [48] and mostly affects the jaw landmarks which are highly dependent on changes in head pose. It should be noted that in our work we do not use any explicit supervision for landmark detection nor focus on ad-hoc methods to cope with this issue. Errors on jaw landmarks due to invisible coordinates and improvements thanks to deformable models can be observed in qualitative results (Fig. 12).

Herein, we evaluate the landmark localization performance of DenseReg as well as the performance obtained by employing DenseReg as an initialization for deformable models [32, 40, 1, 38] trained for the specific task. In the second scenario, we provide a slightly improved initialization with very small computational cost by reconstructing the detected landmarks with a PCA shape model that is con-

structed from ground-truth annotations.

We present experimental results using the very challenging 300W benchmark. This is the testing database that was used in the 300W competition [36, 35] - the most important facial landmark localization challenge. The error is measured using the point-to-point RMS error normalized with the interocular distance and reported in the form of Cumulative Error Distribution (CED). Figure 7 (bottom) presents some self-evaluations in which we compare the quality of initialization for deformable modelling between DenseReg and two other standard face detection techniques (HOG-SVM [23], DPM [30]). The employed deformable models are the popular generative approach of patch-based Active Appearance Models (AAM) [32, 40, 1], as well as the current state-of-the-art approach of Mnemonic Descent Method (MDM) [38]. It is interesting to notice that the performance of DenseReg without any additional deformable model on top, already outperforms even HOG-SVM detection combined with MDM. Especially when DenseReg is combined with MDM, it greatly outperforms all other combinations.

Method	*AUC*	*Failure Rate (%)*
DenseReg + MDM	**0.5219**	**3.67**
DenseReg	0.3605	10.83
Fan et al. [15]	0.4802	14.83
Deng et al. [14]	0.4752	5.5
Martinez et al. [28]	0.3779	16.0
Cech et al. [8]	0.2218	33.83
Uricar et al. [42]	0.2109	32.17

Table 2: Landmark localization results on the 300W testing dataset using 68 points. Accuracy is reported as the Area Under the Curve (AUC) and the Failure Rate of the Cumulative Error Distribution of the RMS point-to-point error normalized with interocular distance.

Figure 7 (top) compares DenseReg+MDM with the results of the latest 300W competition [35], i.e. Cech et al. [8], Deng et al. [14], Fan et al. [15], Martinez et al. [28] and Uricar et al. [42]. We greatly outperform all competitors by a large margin. It should be noted that the participants of the competition did not have any restrictions on the amount of training data employed and some of them are industrial companies (e.g. Fan et al. [15]), which further illustrates the effectiveness of our approach. Finally, Table 2 reports the area under the curve (AUC) of the CED curves, as well as the failure rate for a maximum error of 0.1. Apart from the accuracy improvement shown by the AUC, we believe that the reported failure rate of 3.67% is remarkable and highlights the robustness of DenseReg.

4.4. Deformable Tracking

For the challenging task of deformable face tracking on lengthy videos, we employ the testing database of the

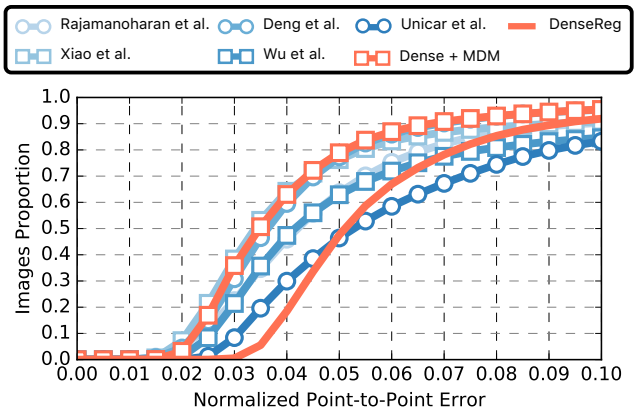


Figure 8: Deformable tracking results against the state-of-the-art on the 300VW testing dataset using 68 points. Accuracy is reported as Cumulative Error Distribution of RMS point-to-point error normalized with interocular distance.

300VW challenge [37, 12] - the only existing benchmark for deformable tracking "in-the-wild". The benchmark consists of 114 videos ($\sim 218k$ frames in total) and includes videos captured in totally arbitrary conditions (severe occlusions and extreme illuminations). The tracking is performed based on sparse landmark points, thus we follow the same strategy as in the case of landmark localization in Sec. 4.3.

We compare the output of DenseReg, as well as DenseReg+MDM which was the best performing combination for landmark localization in static images (Sec. 4.3), against the participants of the 300VW challenge: Yang et al. [45], Uricar et al. [41], Xiao et al. [44], Rajamanoharan et al. [34] and Wu et al. [43]. Figure 8 reports the CED curves for all video categories, whereas Table 3 reports the AUC and Failure Rate measures. DenseReg combined with MDM demonstrates the best performance, even by a short margin from the winner of the 300VW competition. However, it should be highlighted that our approach is not fine-tuned for the task-at-hand as opposed to the rest of the methods that were trained on video sequences and most of them make some kind of temporal modelling. Finally, similar to the 300W case, the participants were allowed to use unlimited training data (apart from the provided training seuqences), as opposed to DenseReg (and MDM) that were trained only on the 3148 images mentioned in Sec. 4.1. Please refer to the supplementary material for a more detailed presentation of the tracking results.

4.5. Monocular Depth Estimation

The fitted template shapes also provide the depth from the image plane. We transfer this information to the visible pixels on the image using the same z-buffering operation used for the deformation-free coordinates (detailed in Sec. 2 of the paper). We adopt this as an additional su-

Method	*AUC*	*Failure Rate (%)*
DenseReg + MDM	**0.5937**	**4.57**
DenseReg	0.4320	8.1
Yang et al. [45]	0.5832	4.66
Xiao et al. [44]	0.5800	9.1
Rajamanoharan et al. [34]	0.5154	9.68
Wu et al. [43]	0.4887	15.39
Unicar et al. [41]	0.4059	16.7

Table 3: Deformable tracking results against the state-of-the-art on the 300VW testing dataset using 68 points. Accuracy is reported as the Area Under the Curve (AUC) and the Failure Rate of the Cumulative Error Distribution of the RMS error normalized with interocular distance.

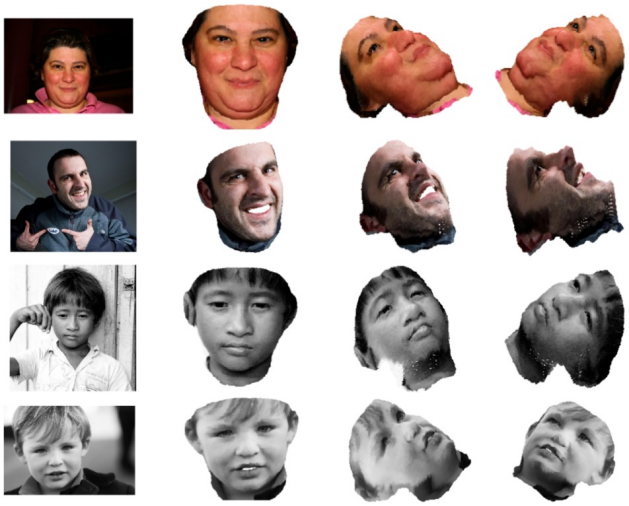


Figure 9: Exemplar 3D renderings obtained using estimated depth values.

pervision signal: $Z \in [0, 1]$ and add another branch to our network to estimate the depth along with the deformation-free coordinates. To our knowledge, there is no existing results in literature that would allow a quantitative comparison. We are providing example reconstructions using estimated monocular depth fields at Fig.9. We observe that this additional branch does not affect the performance of other branches and adds little to the complexity, since it is just a 1x1 convolution layer after the final shared convolutional layer.

4.6. Ear Landmark Localization

In order to highlight the ability of DenseReg to generalize to any kind of deformable object, we report experimental results on the human ear. We employ the 602 images and sparse landmark annotations that were generated in a semi-supervised manner by Zhou et al. [47]. Due to the lack of a 3D model of the human ear, we apply Thin Plate Splines to bring the images in dense correspondence and create the deformation-free space. Then, we perform landmark localization following the same procedure as in Sec. 4.3. We split the images in 500 for training and 102 for testing.

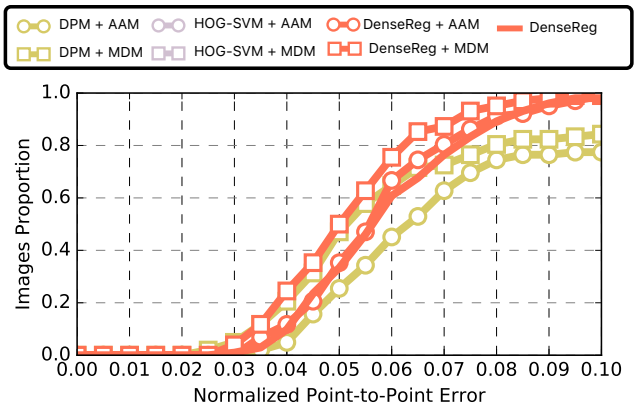


Figure 10: Landmark localization results on human ear using 55 points. Accuracy is reported as Cumulative Error Distribution of normalized RMS point-to-point error.

Given the lack of state-of-the-art deformable models on human ear, we compare DenseReg with DenseReg+AAM and DenseReg+MDM. We also trained a DPM detector in order to compare the initialization quality with DenseReg. Figure 10 reports the CED curves based on the 55 landmark points using the RMS point-to-point error normalized by the bounding box average edge length and on Table.4, we provide failure rate and the Area Under Curve(AUC) measures. Once again, the results are highly accurate even without improving DenseReg with a deformable model.

Method	*AUC*	*Failure Rate (%)*
DenseReg + MDM	**0.4842**	**0.98**
DenseReg	0.4150	1.96
DenseReg + AAM	0.4263	0.98
DPM + MDM	0.4160	15.69
DPM + AAM	0.3283	22.55

Table 4: Landmark localization results on human ear using 55 points. Accuracy is reported as the Area Under the Curve (AUC) and the Failure Rate of the Cumulative Error Distribution of the normalized RMS point-to-point error.

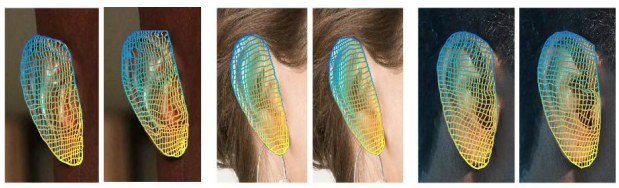


Figure 11: Exemplar pairs of deformation-free coordinates of dense landmarks on human ear. *Left:* Estimated by DenseReg. *Right:* Ground-truth produced by TPS.

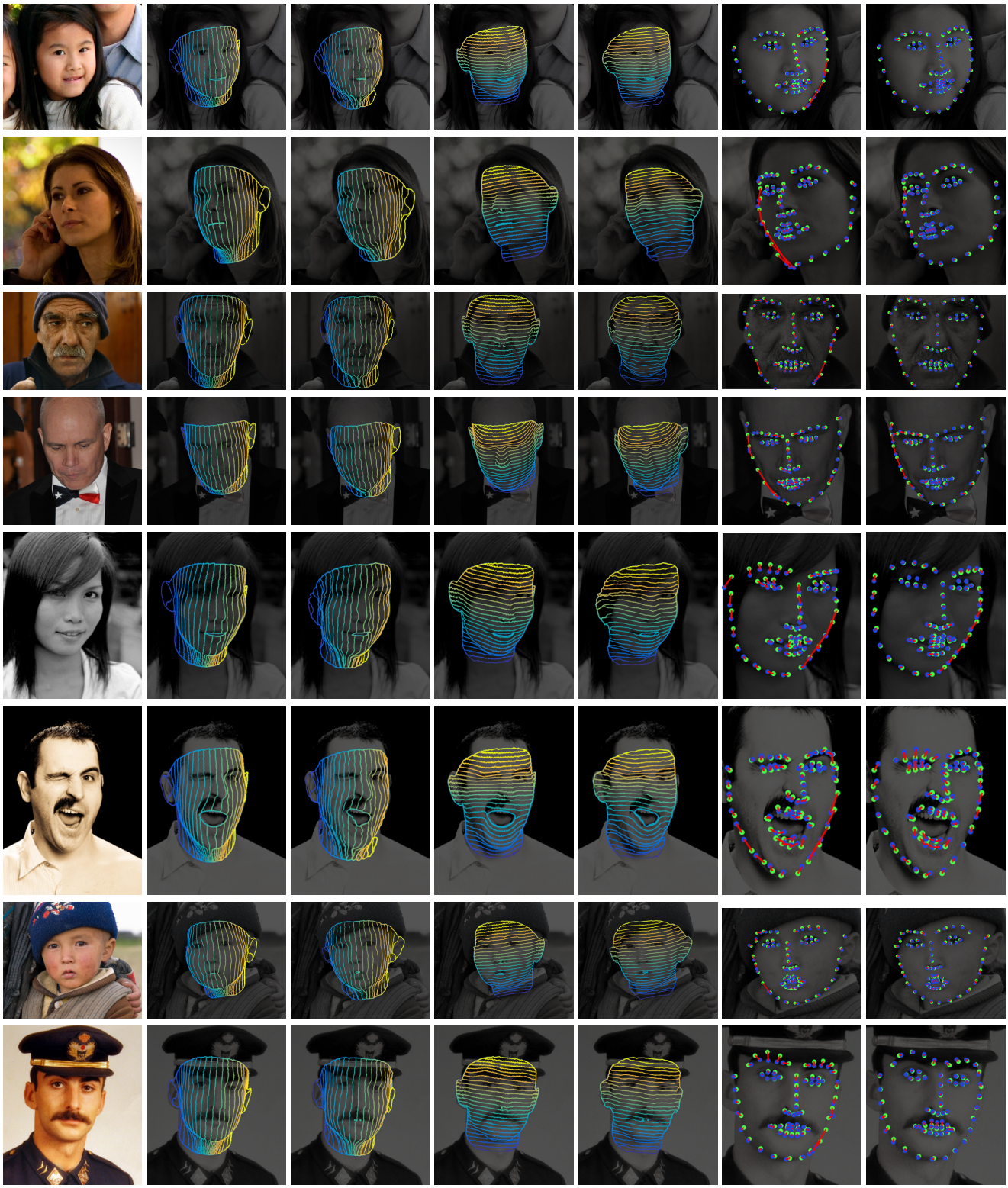


Figure 12: Qualitative Results. From left to right: Original image, ground-truth horizontal coordinates(u^h), estimated horizontal coordinates($\hat{u}^h$) , ground-truth vertical coordinates(u^v), estimated vertical coordinates($\hat{u}^v$) , Landmarks for DenseReg, Landmarks for DenseReg+MDM. Estimated landmarks(blue), ground-truth(green), lines between estimated and ground-truth landmarks(red).

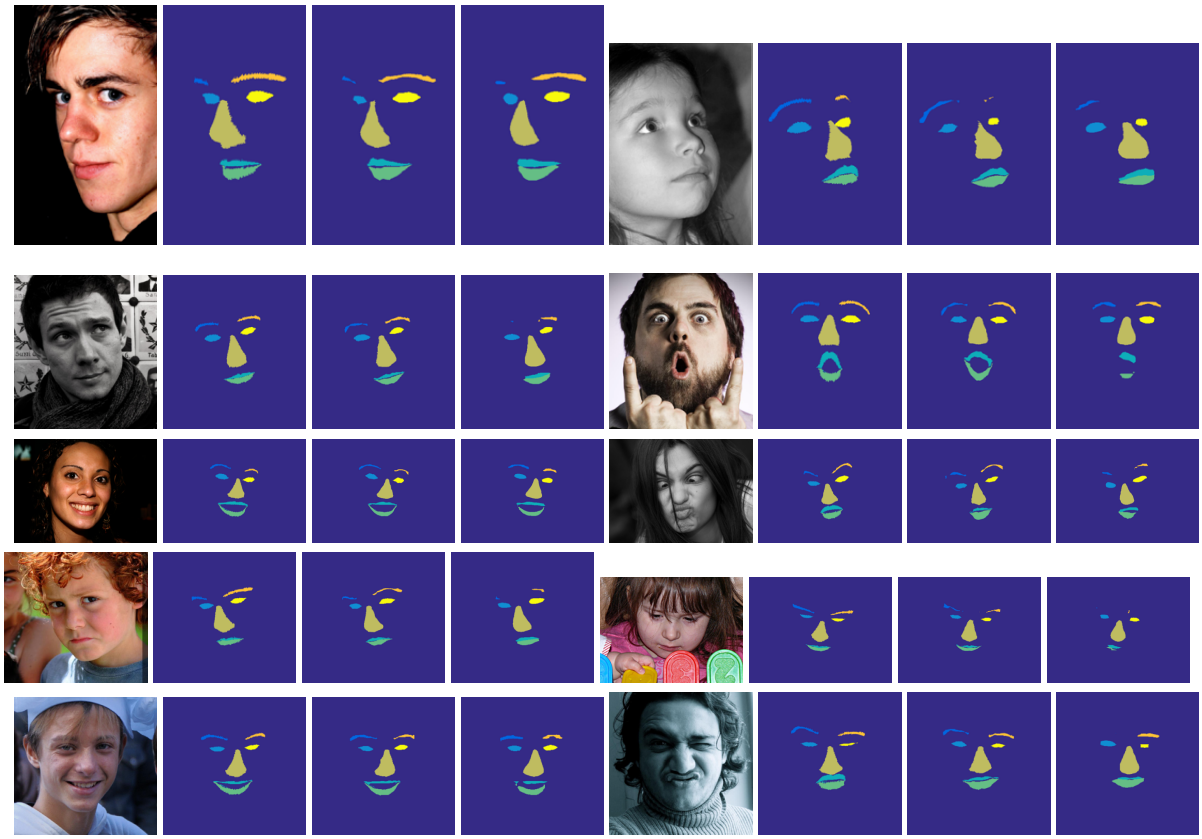


Figure 13: Exemplar semantic segmentation results. *Left:* Ground-truth. *Center:* DenseReg. *Right:* DeepLab-v2.

5. Conclusion

We propose a fully-convolutional regression approach for establishing dense correspondence fields between objects in natural images and three-dimensional object templates. We demonstrate that the correspondence information can successfully be utilised on problems that can be geometrically represented on the template shape. Throughout the paper, we focus on face shapes, where applications are abundant and benchmarks allow a fair comparison. We show that using our dense regression method out-of-the-box outperforms a state-of-the-art semantic segmentation approach for the task of face-part segmentation, while when used as an initialisation for SDMs, we obtain the currently best results on the challenging 300W landmark localization challenge. We believe that our method will find ubiquitous use, since it can be readily used for face-related tasks and can be easily integrated to many other correspondence problems.

References

[1] E. Antonakos, J. Alabort-i-Medina, G. Tzimiropoulos, and S. Zafeiriou. Feature-based lucas-kanade and active appearance models. *IEEE Transactions on Image Processing*, 24(9):2617–2632, September 2015. 6, 7

[2] 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. 5

[3] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway. A 3d morphable model learnt from 10,000 faces. 3

[4] J. Booth and S. Zafeiriou. Optimal uv spaces for facial morphable model construction. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4672–4676. IEEE, 2014. 3

[5] D. Boscaini, J. Masci, E. Rodolà, and M. M. Bronstein. Learning shape correspondence with anisotropic convolutional neural networks. *arXiv preprint arXiv:1605.06437*, 2016. 2

[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. 3

[7] A. Bulat and G. Tzimiropoulos. Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In *European Conference on Computer Vision*, pages 616–624. Springer, 2016. 3

[8] J. Čech, V. Franc, M. Uřičář, and J. Matas. Multi-view facial landmark detection by using a 3d shape model. *Image and Vision Computing*, 47:60–70, 2016. 7

[9] D. Chen, G. Hua, F. Wen, and J. Sun. Supervised transformer network for efficient face detection. In *ECCV*, 2016. 1

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 2, 3, 5

[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*, 2016. 3, 6

[12] G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape. Offline deformable face tracking in arbitrary videos. In *Proceedings of IEEE International Conference on Computer Vision, 300 Videos in the Wild (300-VW): Facial Landmark Tracking in-the-Wild Challenge & Workshop (ICCVW'15)*, Santiago, Chile, December 2015. 7

[13] T. F. Cootes, G. J. Edwards, C. J. Taylor, et al. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001. 3

[14] J. Deng, Q. Liu, J. Yang, and D. Tao. M 3 csr: Multi-view, multi-scale and multi-component cascade shape regression. *Image and Vision Computing*, 47:19–26, 2016. 7

[15] H. Fan and E. Zhou. Approaching human level facial landmark localization by deep learning. *Image and Vision Computing*, 47:27–35, 2016. 7

[16] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 4

[17] U. Grenander, Y. Chow, and D. M. Keenan. *Hands: A Pattern Theoretic Study of Biological Shapes*. Springer-Verlag New York, Inc., New York, NY, USA, 1991. 3

[18] A. Handa, M. Blösch, V. Patraucean, S. Stent, J. McCormac, and A. J. Davison. gvnn: Neural network library for geometric computer vision. *CoRR*, abs/1607.07405, 2016. 1

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 5

[20] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. *CoRR*, abs/1506.02025, 2015. 1

[21] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994. 4

[22] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2, 3

[23] D. E. King. Max-margin object detection. *arXiv preprint arXiv:1502.00046*, 2015. 7

[24] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on computers*, 42(3):300–311, 1993. 1

[25] 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. 5, 6

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 5

[27] S. Mallat. *A wavelet tour of signal processing*. Academic press, 1999. 5

[28] B. Martinez and M. F. Valstar. L 2, 1-based regression and prediction accumulation across views for robust facial landmark detection. *Image and Vision Computing*, 47:36–44, 2016. 7

[29] J. Masci, D. Boscaini, M. Bronstein, and P. Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 37–45, 2015. 2

[30] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision*, pages 720–735. Springer, 2014. 5, 7

[31] G. Papandreou, I. Kokkinos, and P. Savalle. Modeling local and global deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 390–399, 2015. 1

[32] G. Papandreou and P. Maragos. Adaptive and constrained algorithms for inverse compositional active appearance model fitting. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 6, 7

[33] M. Pedersoli, R. Timofte, T. Tuytelaars, and L. Van Gool. An elastic deformation field model for object detection and tracking. *International Journal of Computer Vision*, 111(2):137–152, 2015. 1

[34] G. Rajamanoharan and T. F. Cootes. Multi-view constrained local models for large head angle facial tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 18–25, 2015. 7, 8

[35] 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, 7

[36] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of IEEE Intl Conf. on Computer Vision (ICCV-W 2013), 300 Faces in-the-Wild Challenge (300-W)*, Sydney, Australia, December 2013. 5, 7

[37] J. Shen, S. Zafeiriou, G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of IEEE International Conference on Computer Vision, 300 Videos in the Wild (300-VW): Facial Landmark Tracking in-the-Wild Challenge & Workshop (ICCVW'15)*, December 2015. 7

[38] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR16), Las Vegas, NV, USA*, 2016. 2, 6, 7

[39] S. Tsogkas, I. Kokkinos, G. Papandreou, and A. Vedaldi. Deep learning for semantic part segmentation with high-level guidance. *arXiv preprint arXiv:1505.02438*, 2015. 3

[40] G. Tzimiropoulos and M. Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2014. 6, 7

[41] M. Uricár, V. Franc, and V. Hlavác. Facial landmark tracking by tree-based deformable part model based detector. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 10–17, 2015. 7, 8

[42] M. Uřičář, V. Franc, D. Thomas, A. Sugimoto, and V. Hlaváč. Multi-view facial landmark detector learned by the structured output svm. *Image and Vision Computing*, 47:45–59, 2016. 7

[43] Y. Wu and Q. Ji. Shape augmented regression method for face alignment. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 26–32, 2015. 7, 8

[44] S. Xiao, S. Yan, and A. A. Kassim. Facial landmark detection via progressive initialization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 33–40, 2015. 7, 8

[45] J. Yang, J. Deng, K. Zhang, and Q. Liu. Facial shape tracking via spatio-temporal cascade shape regression. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 41–49, 2015. 7, 8

[46] A. L. Yuille. Deformable templates for face recognition. *Journal of Cognitive Neuroscience*, 3(1):59–70, 1991. 3

[47] Y. Zhou, E. Antonakos, J. Alabort-i Medina, A. Roussos, and S. Zafeiriou. Estimating correspondences of deformable objects "in-the-wild". In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 8

[48] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796, 2015. 6

[49] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012. 5