Deep & Deformable: Convolutional Mixtures of Deformable Part-based Models

Kritaphat Songsri-in[1], George Trigeorgis[1], and Stefanos Zafeiriou[1,2]
[1] Department of Computing, Imperial College London, UK
[2] Center for Machine Vision and Signal Analysis, University of Oulu, Finland
{*kritaphat.songsri-in11, george.trigeorgis08, s.zafeiriou*}*@imperial.ac.uk*

***Abstract*— Deep Convolutional Neural Networks (DCNNs) are currently the method of choice for tasks such that objects and parts detections. Before the advent of DCNNs the method of choice for part detection in a supervised setting (*i.e.*, when part annotations are available) were strongly supervised Deformable Part-based Models (DPMs) on Histograms of Gradients (HoGs) features. Recently, efforts were made to combine the powerful DCNNs features with DPMs which provide an explicit way to model relation between parts. Nevertheless, none of the proposed methodologies provides a unification of DCNNs with strongly supervised DPMs. In this paper, we propose, to the best of our knowledge, the first methodology that jointly trains a strongly supervised DPM and in the same time learns the optimal DCNN features. The proposed methodology not only exploits the relationship between parts but also contains an inherent mechanism for mining of hard-negatives. We demonstrate the power of the proposed approach in facial landmark detection "in-the-wild" where we provide state-of-the-art results for the problem of facial landmark localisation in standard benchmarks such as 300W and 300VW.**

I. INTRODUCTION

Objects and parts detections in unconstrained imagery are challenging problems in the intersection of machine learning and computer vision with many commercial applications. Object detection general refers to the problem of determining generic objects bounding box positions in given images such as faces or humans detections used in the surveillance system [52], [26]. Currently, various state-of-the-art methods in generic objects detections capitalise on the power of Deep Convolutional Neural Networks (DCNNs) to learn features that are approximately invariant to object's deformations [19], [39], [25] and [33].

On the other hand, in the problem of estimating the position of objects parts *e.g.*, facial landmark localisation, it is advantageous to learn and use the relationship between parts. Hence, many popular methods such as strongly supervised Deformable Part-based models (DPMs) [30], Active Appearance Models (AAMs) [6] and Constrained Local Models (CLMs) [7] learn statistical models for both object's appearance and shape (deformations). Traditionally, these methods represented the appearance of the object by hand-crafted features namely Scale-Invariant Feature Transform (SIFT) and Histogram of Gradient (HoG). However, with the availability of a large amount of annotated data, cascade regression methodologies, have started to become quite popular and achieved state-of-the-art results. Arguably the most prominent cases are the Supervised Descent Method proposed in [48], [47] which is based on HoG and the recently introduced end-to-end trainable architecture MDM [40] which combines shallow Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to learn a cascade regression from image pixels to landmark positions. The above-noted methods require an initialisation, generally provided by the application of a face detector. For the case of large pose variations the current state-of-the-art includes the use of very deep neural network architectures, inspired by body pose estimation which first learn store maps for each part and then regress the score maps to part locations [4] (the object's shape is implicitly modelled in the regression step). Nevertheless, such methods have not reported results in the current established benchmarks such as 300W [36], [35] and 300VW [38], [5].

Our idea in this paper is to bridge the gap between DPMs and current DCNNs methods, leading to an end-to-end trainable DPM architecture where the features are learned by fully convolutional deep neural networks. By explicit statistical modelling of the object's deformation, we do not only learn better final responses for face detection and landmark localization but also refine the fully convolutional network that learns the score map for each part. Please see Fig. 1 for an example.

II. RELATED WORK

Before the advent of DCNNs the methods of choice for object detection were variants of DPMs [11], [12]. In original discriminatively trained DPM architecture [11], the part representation did not necessarily correspond to semantically meaningful parts of objects (also referred as weakly supervised DPMs). That is, the parts were discovered automatically using the latent Support Vector Machine (SVM) architecture in order to facilitate object detection. Furthermore, in order to model deformations, a simple star graph was chosen that describes the object in various scales. Nevertheless, soon it was made evident that if the parts were provided by a human annotation process then training a DPM with strong part supervision not only leads to better detection performance, but also much fewer data are required for training a good object detector [30], [51].

Fallen out of fashion due to the rise of SVMs, CNNs has been brought to the attention again in 2012 by Krizhevsky *et al.*[23]. They showed substantially higher image clas-

978-1-5386-2335-0/18/$31.00

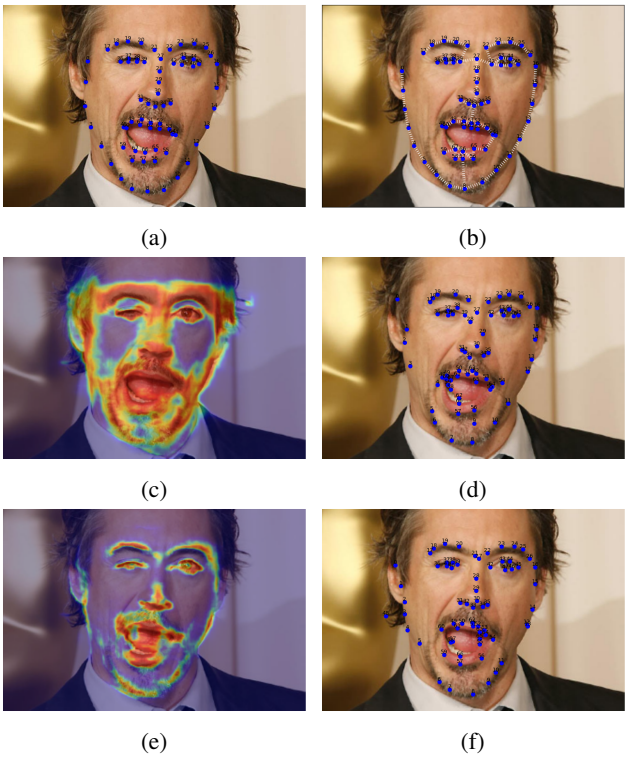


Fig. 1: (a) Ground-truth face landmarks locations. (b) Landmarks obtained from our end-to-end DPMs on top of DCNN heat-map. (c) The appearance model learned from our pre-trained DCNN. (d) The maximum likelihood estimate of the landmarks extracted from the appearance model (c). (e) The appearance model learned from our end-to-end DCNN. (f) The maximum likelihood estimate of the landmarks extracted from the appearance model (e). By incorporating an explicit geometry prior, we can improve on the prediction results (compare (b) to (d) and (f)).

sification accuracy on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [34] by training a large classification DCNN, now known as AlexNet, on 1.2 million labelled images. The current state-of-the-art methods for object detection make use of classification DCNNS and object proposals. Notable examples include R-CNN [16], Fast R-CNN [15], Faster R-CNN [33] and SSD [25]. A link between weakly supervised DPMs and CNNs was made in [17].

One of the first approach to combine DPMs with CNNs was made in [37] (so-called C-DPM) where convolution layers of the AlexNet network were used as a feature extractor in a weakly supervised DPM. They reported that by substituting baseline HoG feature with AlexNet feature, they obtained a substantial boost in performance. However, it still cannot compete with the R-CNN. Another early approach is the so-called DenseNet [20] which provides dense, multi-scale features pyramid from AlexNet that can substitute low-level feature pyramid such as HoG that has been used widely in deformable models. A method that combined DCNN pyramidal features with weakly supervised DPMs was proposed in [32] and achieved state-of-the-art performance for face detection.

An interesting point is raised by Faceness-Net [49] which considered face detection from a different perspective through scoring facial parts responses by their spatial structure and arrangement. They trained independent CNNs for each part of the face (*e.g.*hair, eyes, nose, mouth, beard) to create part response map in which they call "partness" map. A set of candidate windows are then generated (similar to proposing regions in the R-CNN) to be scored based on the combined partness maps. Lastly, the high scored candidates are refined further by being fed to a jointly trained face classification and bounding box regression CNN whose features are based on AlexNet. With this approach, they achieved outstanding performance on face detection challenge in the wild [22] [30]. This gives an insight that facial landmark localisations using DPMs, even by using a heuristic approach, with convolution features may also benefit from using different features for each facial landmarks or even from only sharing features with the neighbour landmarks that look similar *e.g.*. landmarks around eye browns can share the same classification CNN.

Our model is related to [50] but is different in many ways. Compared to our model which use mixtures of components to represent the object such as faces at many viewpoints, they use a single component with the mixture of parts which is more suitable for the task of human pose estimation since each part can have various appearance due to extreme orientations and occlusions. Furthermore, the DCNN in our model also takes into account the problem of objects and parts detections at multiple scales.

Based on the evidence above we propose to develop a mixture of DPMs for jointly solving face detection, pose estimation, and landmark localisation based on part-dependent features that learned by DCNNs. Our approach is end-to-end trainable and learns jointly both the images features filters and DPMs part filters. Contrary to the state-of-the-art approaches, such as MDM, our approach does not need any initialization. Furthermore, it can nicely be combined with MDM and even trained in an end-to-end manner. The combination of our approach with MDM produces state-of-the-art results in the standard benchmarks in facial landmark localisation, such as 300W and 300VW.

III. Model

We use the model of [14] and [30] to combine appearance and deformation terms. That is, objects are represented by a collection of parts, and they are acyclic connected with deformable configuration. Each model's part captures local appearance's properties by producing part's response map. To reduce complexity, their spring-like connections are deliberately modelled by quadratic distances between connecting parts locations. Although the model can express variations in object appearance and shape, mixture models are used to deal with significant object variations such as pose and rotation. Lastly, since the models capture object

and part locations at a single scale, image pyramids are used to determine object at multiple scales.

To describe the models formally, let each tree model be $T_m = (V_m, E_m)$ where m indicates a mixture, and V_m and E_m are tree vertices and connecting edges respectively. For a positive image $\mathbf{I}$, let $\ell_i = (x_i, y_i)$ be a location of part $i \in V_m$. A configuration of parts $\mathcal{L} = \{\ell_i : i \in V_m\}$ can be scored as described below:

$$\mathcal{C}(I, \mathcal{L}, m) = \mathcal{A}_m(\mathbf{I}, \mathcal{L}) - \mathcal{S}_m(\mathcal{L}) + \alpha^m \quad (1)$$

$$\mathcal{A}_m(I, \mathcal{L}) = \sum_{i \in V_m} \mathbf{w}_i^m \cdot \phi^m(\mathbf{I}, \ell_i) \quad (2)$$

$$\mathcal{S}_m(\mathcal{L}) = \sum_{ij \in E_m} (a_{ij}^m, b_{ij}^m, c_{ij}^m, d_{ij}^m) \cdot \bar{\phi}^m(\ell_i, \ell_j) \quad (3)$$

The score $\mathcal{C}(\mathbf{I}, \mathcal{L}, m)$ in (1) is the appearance scores in (2) minus its shape deformation cost in (3) plus bias α^m which determine how likely that an object and its parts will be in an image $\mathbf{I}$ at configuration $\mathcal{L}$.

1) Parts appearance term: The parts appearance term in (2) is the summation over each part between the product of part parameter w_i^m and a patch of extracted feature $\phi^m(\mathbf{I}, \ell_i)$. In [31] and [14], $\phi^m(\mathbf{I})$ is computed by HOG feature for every mixture m. In our model, $\phi^m(\mathbf{I})$ is the features extracted by a DCNN that is based on ResNet-50. The overall DCNN architecture is shown in Fig. 2. We use skip layers [18] that take intermediate layer activations as inputs and perform simple linear operations on those using convolutions. In particular, we pool features from layers conv1, block2/unit$_4$, block3/unit$_6$, block4/unit$_3$ which show up as $C_1, \cdots, C_4$. Modifying [3] slightly, we process these intermediate layers with batch normalization [21] as $B_1, ..., B_4$ to bring the intermediate activations to a common range. To account for the varying face sizes in images we employ a 3-scale pyramid with tied weights of our proposed network where at scales 2 & 3 we down-sample the image by half and quarter times respectively by using a 2D average pooling operation. To aid the learning of the model we add intermediate supervision using our loss function K at each of the 3 scales and also to the final multi-scale fused result.

In [30], the term $w_i^m \cdot \phi^m(\mathbf{I}, \ell_i)$ in (2) is used to produce a response map for each landmark. However, we use our network to produce each landmark response map directly. In particular, the outputs of the network are a stack of each part's response map that has the same size as the given image I with $|V_m| + 1$ channels where $|V_m|$ is the number of landmarks and 1 represents an extra background class. Finally, the softmax function is used to produce the probability of the i^{th} part at location ℓ_i. In our case, the term w_i^m acts as the coefficients that allow the scores of each mixture to be comparable.

2) Deformable configuration term: The deformable terms define the penalty of the distance between the locations of the connecting parts from their rest locations. The parameters $(a_{ij}^m, b_{ij}^m, c_{ij}^m, d_{ij}^m)$ in (3) penalize locations that are further away from their ideal locations v_{ij}^m. Given connecting parts location ℓ_i and ℓ_j, $\bar{\phi}(\ell_i, \ell_j)$ is the quadratic and linear

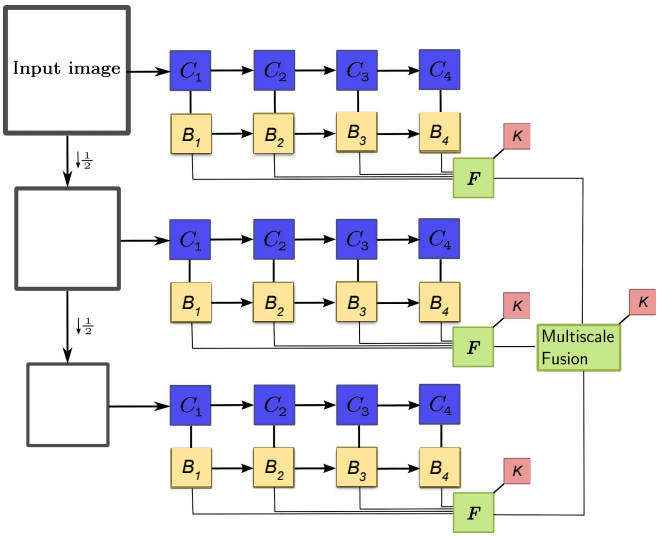


Fig. 2: The proposed multi-scale ResNet-50 based architecture.

distance between ℓ_i and ℓ_j relative to v_i^m defined as:

$$\bar{\phi}(\ell_i, \ell_j) = (dx_{ij}^2, dx_{ij}, dy_{ij}^2, dy_{ij}) \quad (4)$$

where $(dx_{ij}, dy_{ij}) = (x_i, y_i) + v_{ij}^m - (x_j, y_j)$. To keep the model simple, the ideal location v_{ij}^m are fixed and precomputed by averaging the distance between parts locations after procrustes analysis of all positive examples within the same mixture.

3) Bias term: Finally, the bias term α_m is a scalar prior associated with each mixture m. It helps the scores of multiple models to be comparable when we combine them into mixture models.

IV. Inference

Face and landmarks detection in an image $\mathbf{I}$ given model parameters correspond to maximizing the score $\mathcal{C}(\mathbf{I}, \mathcal{L}, m)$ in (1) over all configurations $\mathcal{L}$ and mixtures m:

$$\mathcal{C}^*(\mathbf{I}, \mathcal{L}, m) = \max_m(\max_{\mathcal{L}}(\mathcal{C}(\mathbf{I}, \mathcal{L}, m))) \quad (5)$$

In order to separate positive examples from negative examples, we put a threshold to the above score. Considering the appearance scores in (2) and the deformable cost in (3), it is easy to see that the complexity is dominated by the latter as it required to compute the values at each connecting part location ℓ_i and ℓ_j. Solving this problem naively by iterating through each two possible locations ℓ_i and ℓ_j would take $O(h^2)$ where h denotes the number of possible locations. However, because the model is assumed to be a tree and the deformable costs are in quadratic terms, this problem can be solved by using dynamic programming and generalized distance transforms [13] which is very efficient and can reduce the complexity to be $O(h)$. Hence, the overall complexity of solving the inference problem is $O(M|V_m|h)$ where M is the number of mixtures and $|V_m|$ is the number of parts in each mixture.

V. Model Learning

To put in the simplest form, our end-to-end model is to put together the DCNN that produce face landmarks response with DPMs that learn to classify mixture of face examples from negative examples. Since multiple faces "in the wild" datasets are annotated with part locations, but without pose, we define the mixture by clustering the procrustes analysis of landmarks annotations.

A. Pretraining DCNN

We use a ResNet-50 based DCNN to produce each object part response feature with $|V_m| + 1$ channels that are up-sampled to the same size as the given image. For a given positive image $\mathbf{I}$ and its landmark's locations $\mathcal{L} = \{\ell_i : i \in V_m\}$, we define ground-truth key-points K at each part channel i as:

$$K(\mathbf{I}, i, \ell_i) = \begin{cases} 1, & \text{if } ||\ell_i - \widetilde{\ell_i}|| \leq \delta \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $\widetilde{\ell_i}$ is the ground truth landmark location and δ is a constant scalar threshold. During pretraining the DCNN, we conducted multiple experiments with different values of δ varied from 1-10 (pixel). We found that 4 resulted in the best performance, and we kept it constant throughout our end-to-end training. The ground-truth key-points have value 1 inside the box, and value 0 outside the box at the corresponding part channels. On the other hand, the ground-truth key-points K are all 0 for negative examples.

Let $\phi^m(\mathbf{I})$ be the DCNN feature that fused response maps at multiple scales of the pyramid and let $K(\mathbf{I})$ be the stack ground-truth key-points of each part, we use discrete cross-entropy $H_{\mathbf{I}}(p, q) = -\sum_L p(\mathbf{I}, L) \log q(\mathbf{I}, L)$ as the loss function between the predicted features and the ground-truth key-points. In addition, we assist the training by also adding intermediate supervision using the same loss function at each scale of the pyramid. Hence, the final loss function is:

$$\sum_{\mathbf{I}} (H_{\mathbf{I}}(K, \phi^m) + \sum_s H_{\mathbf{I}}(K, \phi^m_s)) \tag{7}$$

where $\phi^m_s(\mathbf{I})$ is the intermediate feature at each scale s.

By employing a 3-scale pyramid, the network can be trained with face images at multiple scales. Initially, we aid the training by only feeding images with a single face that are cropped proportionally to the face size. Later, the network is trained with images of faces from the dataset directly. For simplicity, we do not distinguish features of different mixtures at this stage (*i.e.*$M = 1$).

B. End-to-end Training

Traditionally, the DPMs are trained discriminatively with Latent-SVM [14] or MI-SVM [1]. Given labelled positive examples $\{\mathbf{I}, \mathcal{L}, y = 1\}$ and negative examples $\{\mathbf{I}, y = -1\}$, learning parameters is formulated by the hinge-loss as follows:

$$\frac{1}{2}\beta \cdot \beta + C \sum_{\mathbf{I}} \max(0, 1 - y(\beta^m \cdot \psi^m(\mathbf{I}, \mathcal{L}))) \tag{8}$$

where $\beta^m = (..., w^m_i, ..., (a^m_{ij}, b^m_{ij}, c^m_{ij}, d^m_{ij}), ..., \alpha^m)$ is the concatenation of all model parameters and $\psi^m(\mathbf{I}, \mathcal{L}) = (..., \phi^m(\mathbf{I}, \ell_i), ..., \bar{\phi}^m(\ell_i, \ell_j), ..., 1)$ is the concatenation of the corresponding features. The configuration $\mathcal{L}$ for each negative example is the configuration with the highest score computed by inferencing with current model parameters. The term $\beta^m \cdot \psi^m(\mathbf{I}, \mathcal{L})$ is in fact the score $\mathcal{C}(\mathbf{I}, \mathcal{L}, m)$ defined in (1). The hinge-loss states that the scores of all positive examples should be greater than 1 while the score of negative examples should be lower than -1.

1) Adding DCNN loss: By using only the loss function in (8) for end-to-end training, we found that although the model may improve face detections, the learned features are unsuitable for the task of parts detection. As can be seen in Fig. 3b, the response map of the features learned end-to-end with (8) are more scattered around the ground-truth landmarks compared to the pre-trained network that uses the loss function (7) in Fig. 3a. In order to help the model solve both problems of face and landmark detection, we use the combination of loss functions from (7) and (8). With the combined loss functions, the response map of the extracted feature is improved significantly as shown in Fig. 3c.

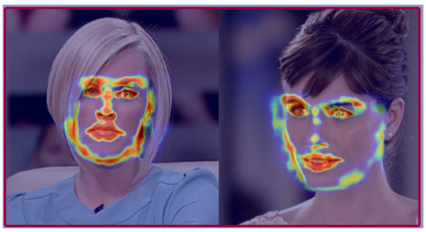
2) Multi-class classification loss: Having a mixture of models, the problem of faces and parts detections are actually classification problem between each mixture and the negative class. However, traditionally DPMs were trained as binary classification problem for each mixture independently. During inference time, the mixture with the highest score is chosen as the predicted class. Interestingly, once we run the experiments with our mixture models, we found that end-to-end training does affect the magnitude of positive examples scores of each mixture differently. Specifically, some mixtures tend to give scores that are in order of magnitude higher than other mixtures even though the given example does not belong to the mixture. We found that mixtures with a fewer number of examples tend to give scores to positive examples relatively higher than the other mixtures. As a result, this behaviour worsens the accuracy for parts detections. Since this issue did not occur in the past for [14] and [51], we argue that when the models are trained end-to-end, the DCNN feature enable the DPMs to discriminate the positive and negative examples significantly better than HoG. As a result, the distance between the support vectors of positive and negative classes for each mixture is further away, creating an ambiguity between mixtures.

In order to also solve the aforementioned ambiguity problem, we propose to use the multi-class hinge loss that also recognizes negative class. Given labelled positive examples $\{\mathbf{I}, \mathcal{L}, \tilde{m}\}$ where $\tilde{m}$ is a true mixture and negative examples $\{\mathbf{I}\}$, we define the loss function as:

$$\frac{1}{2}\beta \cdot \beta + C \sum_{\mathbf{I}} \sum_m \max(0, \tilde{\mathcal{C}}_m) \tag{9}$$

(a) The pre-trained model learned using the loss function (7).

(b) The end-to-end model learned using the loss function (8).



(c) The end-to-end model learned using the combined loss functions of (7) + (8).

Fig. 3: The response map of our ResNet-50 based DCNNs that are produced with different loss functions.

where

$$\mathcal{C}_m = \beta^m \cdot \psi^m(\mathbf{I}, \mathcal{L}) \tag{10}$$

$$\tilde{\mathcal{C}}_m = \begin{cases} 1 - (\mathcal{C}_{\tilde{m}} - \mathcal{C}_m), & \text{if } \mathbf{I} \text{ is positive and } m \neq \tilde{m} \\ 1 - \mathcal{C}_{\tilde{m}}, & \text{if } \mathbf{I} \text{ is positive and } m = \tilde{m} \\ 1 + \mathcal{C}_m, & \text{otherwise} \end{cases} \tag{11}$$

On top of penalising positive examples that score less than 1 and negative examples that score greater than -1, inspired by [44] this loss function also gives the penalty when the scores of the given positive examples with true mixture are not greater than the scores of other mixtures by at least 1. With (9), the inference problem still remains the same as described in (5), but the relationship between each mixture's score are also considered as part of the loss function.

C. Training Algorithm

Our end-to-end model is jointly trained with DCNN loss (7) + multi-class hinge loss (9). During training with stochastic gradient descent, we compute the gradient of (7) and sub-gradient of (9). With back propagation, DPM parameters are updated w.r.t. (9) while DCNN parameters are updated w.r.t (7) + (9). The end-to-end training algorithm is fully described in Algorithm 1.

The function dcnn-loss$(\mathbf{I}_i, K_i)$ compute DCNN loss in (7). The function detect-best$(m, \mathbf{I}_i)$ find landmarks with the highest score described in Section IV. The function dpm-score$(m, \mathcal{L})$ compute DPM score in (1). The function dpm-loss(score) compute multi-class hinge loss in (9). Lastly the function Adam$(\frac{1}{i}\sum_{i=1}^{b} \text{TotalLoss}_i)$ trains β with adam optimizer.

In a similar manner, the model can be easily combined, in an end-to-end fashion, with methods that improve the accuracy further (e.g., MDM).

Data: $(\mathbf{I}_1, \mathcal{L}_1, K_1, m_1), ..., (\mathbf{I}_n, \mathcal{L}_n, K_n, m_n)$
Parameters: β
Result: new parameters β
for *number of epoch* **do**
  for $i \leftarrow 1$ **to** b **do**
    $\text{DCNNLoss}_i \leftarrow$ dcnn-loss$(\mathbf{I}_i, K_i)$ Scores $\leftarrow [\,]$
    for $m \leftarrow 0$ **to** M **do**
      if $m_i == -1$ *or* $m_i \neq m$ **then**
        L $\leftarrow$ detect-best$(m, \mathbf{I}_i)$
      else
        L $\leftarrow \mathcal{L}_i$
      end
      Score $\leftarrow$ dpm-score$(m, \mathcal{L})$
      Append Score to Scores
    end
    $\text{DPMLoss}_i \leftarrow$ dpm-loss$(Scores, m_i)$
    $\text{TotalLoss}_i \leftarrow \text{DCNNLoss}_i + \text{DPMLoss}_i$
  end
  $\beta \leftarrow \text{Adam}(\frac{1}{i}\sum_{i=1}^{b} \text{TotalLoss}_i)$
end

Algorithm 1: End-to-End training algorithm

VI. Experimental Results

To demonstrate the improvement of our model, we focus our experiments on the task of facial landmark detection and tracking on "in-the-wild" datasets 300W [36], [35] and 300VW [38], [5] respectively. Throughout our experiments, we used only 5 mixture components as it already covers most of common pose variations and achieve good results.

A. Training Databases

We train our model with the 300W database annotated with 68 landmarks. The training set consists of the LFPW train set [2], Helen train set [24] and AFW [30] and are combined into 3148 images. The databases contain images of faces under large variations such as pose, occlusion, expression, illumination, age, *etc.*. The negative examples are 1218 images from the INRIAPerson database [8] which tend to be outdoor scenes that do not contain people.

B. In-house baseline

In addition to comparing our results with other state-of-the-art systems, we evaluate various versions of our approaches. We define an out-of-the-box DPM model in [30] as HOG+DPM*. We define RES to be the maximum likelihood estimate of our pre-trained DCNN response map. Similarly, we define RES* to be the maximum likelihood estimate of our end-to-end model's DCNN response map. We define RES+DPM* to be a DPM model that trained on the fixed pre-trained DCNN feature. We define RES*+DPM* to be our end-to-end models trained with the binary hinge loss defined

in (8). Lastly, we define RES*+MUL_DPM* as our end-to-end models trained with the multi-class hinge loss defined in (9).

C. Landmark Localisation on Static Images

We present experimental results on static images using the very challenging 300W benchmark. 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). We evaluated our models with two setting between 68 landmarks and 51 landmarks where jaw line's predictions are ignored.

1) Self evaluation: Fig. 4 show our self-evaluations with 51 (top) and 68 (bottom) landmarks. Both results show a similar trend between the performance of each model. There are couple interesting points to notice from the results. Firstly, all of our models outperform the base-line DPMs model (HOG+DPM*) in [30]. In addition, compared to our pre-trained DCNN, our end-to-end training also improve the maximum likelihood estimate of the DCNN by a big margin. As we mentioned in Subsection V-B, the performance of end-to-end training using loss function (8) (RES*+DPM*) is actually decreased to be roughly at the same level as the maximum likelihood estimate of our pre-trained DCNN (RES). Our best in-house model is RES*+MUL_DPM* which is trained end-to-end with loss function (7) + (9). Please note that our model (RES*+MUL_DPM*) already produces competitive results with the state-of-the-art model MDM whose initialisations are taken from DPMs [28]. Nevertheless, we also combine our model with MDM by utilising our predictions as its initialisations and obtain state-of-the-art results (RES*+MUL_DPM*+MDM).

2) Comparison with State-of-the-art: Fig. 5 compares our best in-house model (RES*+MUL_DPM*) and our combined model (RES*+MUL_DPM*+MDM) with the results of the latest 300W competition [35], *i.e.*Cech *et al.*[43], Deng *et al.*[9], Fan *et al.*[10], Martinez *et al.*[27] and Uricar *et al.*[42]. Our combined method (RES*+MUL_DPM*+MDM) outperforms all competitors. Besides, it should be noted that the participants of the competition did not have any restrictions on the amount of training data and we only used 3148 positive examples. Finally, Table I reports the area under the curve (AUC) of the CED curves, as well as the failure rate (FR) for a maximum error of 0.1. Our combined model (RES*+MUL_DPM*+MDM) achieved remarkable AUC and FR for 68 landmarks at **60.52** and **3.67**% respectively.

D. Landmark Tracking on Videos

For the task of face landmarks tracking, we evaluate our models with the challenging database of the 300VW challenge [38], [5]. The benchmark consists of 114 videos ($\sim$ 218k frames in total) and includes videos captured in totally arbitrary conditions (*e.g.*severe occlusions and extreme pose). The database is separated into 3 categories, each indicates different levels of difficulties. In order to show the robustness of our model, we choose to show the results in category 3 which is the most difficult category [5]. To test our model's limit further, we perform the tracking by only applying face and landmark detection on each frame independently without any pre-initialisation nor failure checking between the frames.

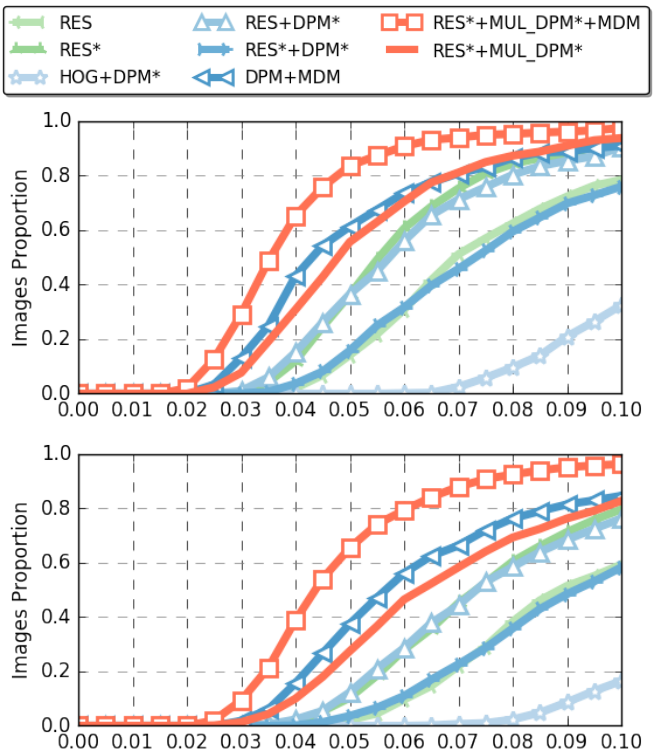


Fig. 4: Our in-house Landmark localization results on the 300W testing dataset with 51 (top) and 68 (bottom) points. The results are reported as Cumulative Error Distribution of RMS point-to-point error normalized with interocular distance.

We compare our models (RES*+MUL_DPM*) and (RES*+MUL_DPM*+MDM) against the participants of the 300VW challenge: Deng *et al.*[9], Unicar *et al.*[41], Xiao *et al.*[46], Rajamanoharan *et al.*[29], and Wu *et al.*[45]. Fig. 6 shows the CED curves for video in category 3 while Table II reports the corresponding AUC and FR measures. Our combined model (RES*+MUL_DPM*+MDM) is among the state-of-the-art results and achieves the fewest failure rate at **9.61**%. However, it should be highlighted that the participants were allowed to use more training data. Besides, our model does not require any initialization nor make use of temporal modelling opposed to the rest of the methods.

VII. CONCLUSIONS AND FUTURE WORKS

A. Conclusions

We propose to bridge the gap between mixtures of DPMs and current DCNN methods, leading to an end-to-end trainable DPM architecture where the features are learned by fully convolutional deep neural networks. In particular, our model is the deep extension of the strongly-supervised DPMs for face detection, pose estimation, and landmark localisation

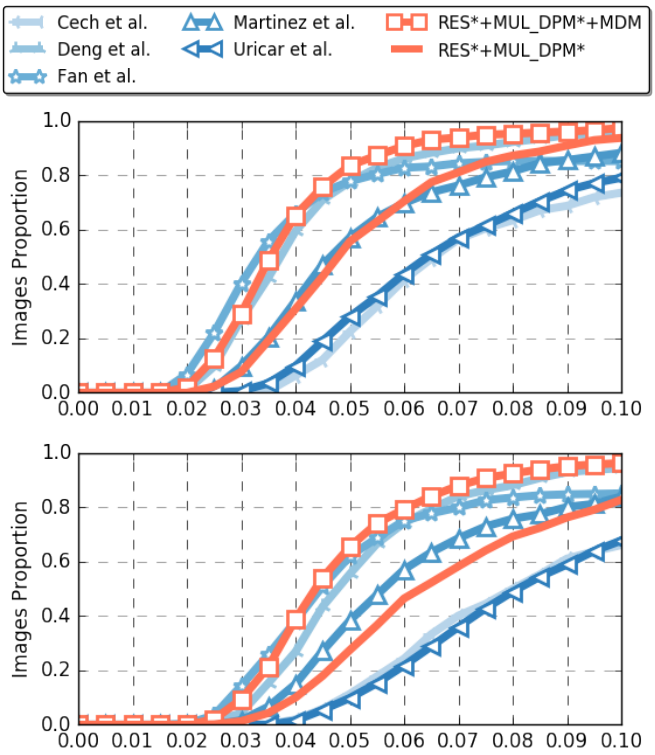


Fig. 5: Our state-of-the-art comparison Landmark localization results on the 300W testing dataset with 51 (top) and 68 (bottom) points. The results are reported as Cumulative Error Distribution of RMS point-to-point error normalized with interocular distance.

[30]. By explicit statistical modelling of the object's deformation, we do not only learn better final responses for face detection and landmark localization but also refine the fully convolutional network that learns the score maps for each part.

B. Future Works

Contrary to other state-of-the-art approaches, our approach does not need any initialization. Although our model can be combined with MDM and produces state-of-the-art results in the standard benchmarks in facial landmark localisation 300W and 300VW, the combined approach can even be trained in an end-to-end manner as an interesting future work.

VIII. Acknowledgements

K. Songsri-in was supported by Royal Thai Government Scholarship. G. Trigeorgis was supported by EPSRC DTA award at Imperial College London and Google Fellowship. The work of S. Zafeiriou was partially funded by EPSRC Project EP/N007743/1 (FACER2VM).

TABLE I: Landmark localization results on the 300W testing dataset using 51 (left) and 68 (right) points. Accuracy is reported as the Area Under the Curve (AUC) and the Failure Rate (FR) of the Cumulative Error Distribution of the RMS point-to-point error normalized with interocular distance

	51 points		68 points	
Method	AUC	FR (%)	AUC	FR (%)
Cech *et al.*[43]	29.51	26.33	22.18	33.83
Deng *et al.*[9]	57.46	3.83	47.52	5.50
Fan *et al.*[10]	57.11	14.67	48.02	14.83
Martinez *et al.*[27]	45.80	11.67	37.79	16.00
Uricar *et al.*[42]	31.86	20.83	21.09	32.17
DPM+MDM [40]	56.34	4.20	45.32	6.80
Our	47.11	6.17	32.86	17.17
Our+MDM	**60.52**	**2.50**	**51.71**	**3.67**

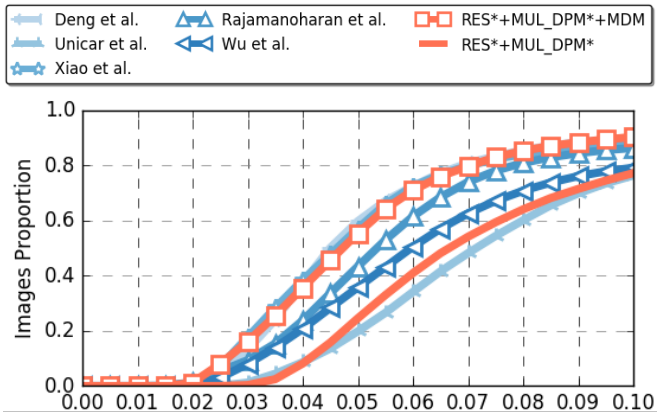


Fig. 6: Deformable tracking results against the state-of-the-art on the 300VW testing dataset using 68 points. Accuracy is reported as Cumulative Error Distribution of RMS point-to-point error normalized with interocular distance

References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems 15*, pages 561–568. MIT Press, 2003.

[2] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 545–552, Washington, DC, USA, 2011. IEEE Computer Society.

[3] S. Bell, C. L. Zitnick, K. Bala, and R. B. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. *CoRR*, abs/1512.04143, 2015.

[4] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. *CoRR*, abs/1609.01743, 2016.

[5] G. G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape. Offline deformable face tracking in arbitrary videos. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):681–685, June 2001.

[7] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *Proceedings of the British Machine Vision Conference*, pages 95.1–95.10. BMVA Press, 2006. doi:10.5244/C.20.95.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society.

[9] J. Deng, Q. Liu, J. Yang, and D. Tao. M^3 CSR: multi-view, multi-scale and multi-component cascade shape regression. *Image Vision Comput.*, 47:19–26, 2016.

[10] H. Fan and E. Zhou. Approaching human level facial landmark

TABLE II: Landmark localization results on category 3 of the 300VW testing dataset using 68 points. Accuracy is reported as the Area Under the Curve (AUC) and the Failure Rate (FR) of the Cumulative Error Distribution of the RMS point-to-point error normalized with interocular distance

Method	AUC	FR (%)
Deng *et al.*[9]	**48.43**	9.88
Unicar *et al.*[41]	28.01	23.57
Xiao *et al.*[46]	48.15	11.92
Rajamanoharan *et al.*[29]	42.28	13.59
Wu *et al.*[45]	36.45	20.78
Our	30.20	22.73
Our + MDM	47.85	**9.61**

localization by deep learning. *Image Vision Comput.*, 47(C):27–35, Mar. 2016.

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[12] P. F. Felzenszwalb, R. B. Girshick, and D. A. McAllester. Cascade object detection with deformable part models. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 2241–2248, 2010.

[13] P. F. Felzenszwalb and D. P. Huttenlocher. Distance transforms of sampled functions. Technical report, Cornell Computing and Information Science, 2004.

[14] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comput. Vision*, 61(1):55–79, Jan. 2005.

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

[16] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587, 2014.

[17] R. B. Girshick, F. N. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks. *CoRR*, abs/1409.5403, 2014.

[18] B. Hariharan, P. A. Arbeláez, R. B. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. *CoRR*, abs/1411.5752, 2014.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[20] F. N. Iandola, M. W. Moskewicz, S. Karayev, R. B. Girshick, T. Darrell, and K. Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *CoRR*, abs/1404.1869, 2014.

[21] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[22] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[24] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part III*, ECCV'12, pages 679–692, Berlin, Heidelberg, 2012. Springer-Verlag.

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. E. Reed. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015.

[26] Y. Lu, J. Zhou, and S. Yu. A survey of face detection, extraction and recognition. *Computers and Artificial Intelligence*, 22(2):163–195, 2003.

[27] B. Martinez and M. F. Valstar. L2,1-based regression and prediction accumulation across views for robust facial landmark detection. *Image Vision Comput.*, 47(C):36–44, Mar. 2016.

[28] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision (ECCV)*, September 2014.

[29] 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.

[30] D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 2879–2886, Washington, DC, USA, 2012. IEEE Computer Society.

[31] D. Ramanan. Dual coordinate solvers for large-scale structural svms. *CoRR*, abs/1312.1743, 2013.

[32] R. Ranjan, V. M. Patel, and R. Chellappa. A deep pyramid deformable part model for face detection. *CoRR*, abs/1508.04389, 2015.

[33] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[35] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: database and results. *Image Vision Comput.*, 47:3–18, 2016.

[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.

[37] P.-A. Savalle, S. Tsogkas, G. Papandreou, and I. Kokkinos. Deformable Part Models with CNN Features. In *European Conference on Computer Vision, Parts and Attributes Workshop*, Zurich, Switzerland, Sept. 2014.

[38] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *2015 IEEE International Conference on Computer Vision Workshop, ICCV Workshops 2015, Santiago, Chile, December 7-13, 2015*, pages 1003–1011, 2015.

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[40] G. Trigeorgis, P. Snape, E. Antonakos, M. A. Nicolaou, and S. Zafeiriou. Mnemonic Descent Method. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[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.

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

[43] J. Čech, V. Franc, M. Uřičář, and J. Matas. Multi-view facial landmark detection by using a 3d shape model. *Image Vision Comput.*, 47(C):60–70, Mar. 2016.

[44] J. Weston and C. Watkins. Multi-class support vector machines. Technical report, Citeseer, 1998.

[45] 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.

[46] S. Xiao, S. Yan, and A. A. Kassim. Facial landmark detection via progressive initialization. In *ICCV Workshops*, pages 986–993. IEEE, 2015.

[47] X. Xiong and F. De la Torre Frade. Supervised descent method and its applications to face alignment. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, May 2013.

[48] X. Xiong and F. D. la Torre. Global supervised descent method. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2664–2673, June 2015.

[49] S. Yang, P. Luo, C. C. Loy, and X. Tang. From facial parts responses to face detection: A deep learning approach. *CoRR*, abs/1509.06451, 2015.

[50] W. Yang, W. Ouyang, H. Li, and X. Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *CVPR*, 2016.

[51] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1385–1392, Washington, DC, USA, 2011. IEEE Computer Society.

[52] S. Zafeiriou, C. Zhang, and Z. Zhang. A survey on face detection in the wild. *Comput. Vis. Image Underst.*, 138(C):1–24, Sept. 2015.