Convolutional Experts Constrained Local Model for 3D Facial Landmark Detection

Amir Zadeh, Yao Chong Lim, Tadas Baltrušaitis, Louis-Philippe Morency
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213, USA
{abagherz,yaochonl,tbaltrus,morency}@cs.cmu.edu

Abstract

Constrained Local Models (CLMs) are a well-established family of methods for facial landmark detection. CE-CLM, the newest member of CLMs, brings CLMs back to state of the art performance. This is done through CE-CLMs ability to model the very complex individual landmark appearance that is affected by expression, illumination, facial hair, makeup, and accessories. A crucial component of CE-CLM is a novel local detector – Convolutional Experts Network (CEN) – that brings together the advantages of neural architectures and mixtures of experts in an end-to-end framework. In this paper we use CE-CLM to learn position of dense 84 landmark positions. To achieve best performance on the Menpo3D dense landmark detection challenge, we use two complementary networks alongside CE-CLM: a network that maps the output of CE-CLM to 84 landmarks called Adjustment Network, and a Deep Residual Network called Correction Networks that learns dataset specific corrections for CE-CLM.

1. Introduction

Facial landmark detection is an essential initial step for a number of research areas such as facial expression analysis, face 3D modeling, facial attribute analysis, emotion recognition, sentiment analysis and person identification [11, 32, 21]. It is a well-researched problem with large amounts of annotated data and has seen a surge of interest in the past couple of years.

One of the most popular methods for facial landmark detection has been the family of Constrained Local Models (CLM) [11, 25]. They model the appearance of each facial landmark individually using *local* detectors and use a shape model to perform *constrained* optimization.

In this paper we use a local detector called Convolutional Experts Network (CEN) that brings together the advantages of neural architectures and mixtures of experts in an end-to-end framework. CEN is able to learn a mixture of experts that capture different appearance prototypes without the need of explicit attribute labeling. For a full facial landmark detection pipeline we use Convolutional Experts Constrained Local Model (CE-CLM)[31], which is a CLM model that uses CEN as a local detector. Our approach is able to extract

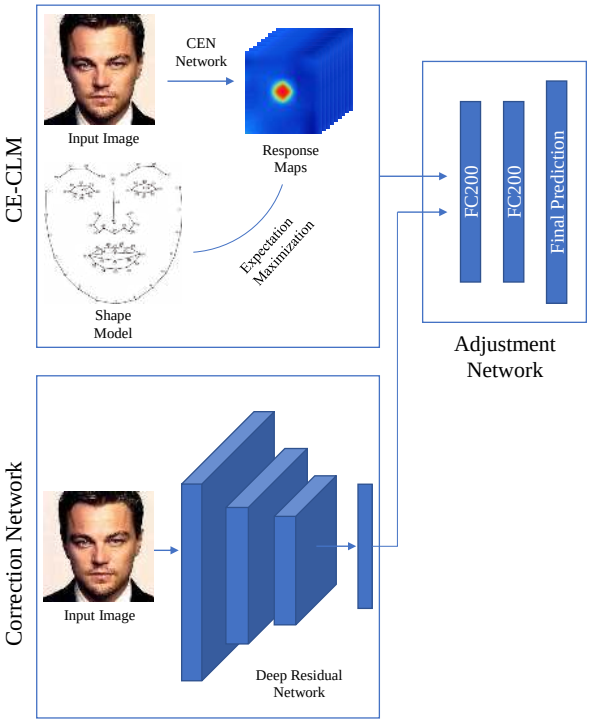


Figure 1: Overview of our pipeline for dense 84-points landmark detection using Convolutional Experts Constrained Local Model. Two networks are used to adapt CE-CLM to 84-points landmark detection: a network that maps the output of CE-CLM to 84 landmarks called Adjustment Network, and a Deep Residual Network called Correction Networks that learns dataset specific corrections for CE-CLM.

3D facial points without being trained on 3D data. This allows our model to be easily adapted for 3D landmark detection. To transfer 68 landmarks to 84, we use a deep neural network network called Adjustment Network. Furthermore, we use a Deep Residual Network called Correction Network to learn the dataset specific corrections to achieve superior results.

Convolutional Experts Constrained Local Model has many qualities that other approaches lack: 1) modeling the appearance of each landmark individually makes Convolutional Experts Constrained Local Model robust to occlusion; 2) Convolutional Experts Constrained Local Model pipeline extracts 3D landmark positions by nature without being trained on 3D data; 3) natural extension to a 3D shape model and multi-view local detectors allow CLMs to deal naturally with pose variations. Thus our model uses the same pipeline to align landmarks to faces without prior information about head pose (no need for frontal vs profile classification); 4) the Expectation Maximization-based model leads to smoothness of tracking in videos.

We evaluate both the benefits of our CEN local detector and then CE-CLM facial landmark detection algorithm through an extensive set of experiments on two publicly-available datasets: 300-W [22], and Menpo 2D Challenge [36]. We finally report our results for Menpo3D dense 84-point landmark detection challenge [6, 35].

2. Related Work

Facial landmark detection plays a crucial role in a number of research areas and applications such as facial attribute detection [18], facial expression analysis [20], emotion recognition and sentiment analysis [33].

Modern facial landmark detection approaches can be split into two categories: *model* and *regression* based. *Model-based* approaches often model both appearance and shape of facial landmarks explicitly with the latter constraining the search space and providing a form of regularization. *Regression-based* approaches on the other hand do not require an explicit shape model and landmark detection is directly performed on appearance.

Model-Based approaches find the best parameters of a face model that match the appearance of an image. A popular model-based method is the Constrained Local Model [11, 25] and its various extensions such as Constrained Local Neural Fields [2, 15] and Discriminative Response Map Fitting [1] which use more advanced methods of computing local response maps and inferring the landmark locations.

Another noteworthy model-based approach is the mixture of trees model [43] which uses a tree based deformable parts model to jointly perform face detection, pose estimation and facial landmark detection. A more recently-proposed 3D Dense Face Alignment method [42] updates the parameters of a 3D Morphable Model [5] using a CNN and has shown good performance on facial landmark detection of profile faces.

Regression-based models predict the facial landmark locations directly from appearance. Majority of such approaches follow a cascaded regression framework, where the landmark detection is continually improved by applying a regressor on appearance given the current landmark estimate in explicit shape regression [7]. Cascaded regression approaches include the Stochastic Descent Method (SDM) [30] and Coarse-to-Fine Shape Searching (CFSS) [41] which attempts to avoid a local optima by performing a coarse to fine shape search. Project out Cascaded regression (PO-CR) [29] is another cascaded regression example that updates the shape model parameters rather than predicting landmark locations directly.

Recent work has also used deep learning techniques for landmark detection. Coarse-to-Fine Auto-encoder Networks [37] use visual features extracted by an auto-encoder together with linear regression. Sun et al. [26] proposed a CNN based cascaded regression approach for sparse landmark detection. Similarly, Zhang et al. [40] proposed to use a CNN in multi-task learning framework to improve facial landmark performance by training a network to also learn facial attributes. Finally, Trigeorgis et al. [28] proposed Mnemonic Descent Method which uses a Recurrent Neural Network to perform cascaded regression on CNN based visual features extracted around landmark locations.

3. Convolutional Experts CLM

Convolutional Experts Constrained Local Model (CE-CLM) algorithm consists of two main parts: response map computation using Convolutional Experts Network (CEN) and shape parameter update using a Point Distribution Model. During the first step, individual landmark alignment is estimated independently of the position of other landmarks. During the parameter update, the position of all landmarks is updated jointly penalizing misaligned landmarks and irregular shapes using a point distribution model. We optimize the following objective:

$$\mathbf{p}^* = \underset{\mathbf{p}}{\operatorname{argmin}} \Big[\sum_{i=1}^{n} -\mathcal{D}_i(x_i; \mathcal{I}) + \mathcal{R}(\mathbf{p}) \Big] \qquad (1)$$

above, $\mathbf{p}^*$ is the optimal set of parameters controlling the position of landmarks (see Equation 3) with $\mathbf{p}$ being the current estimate. $\mathcal{D}_i$ is the alignment probability of landmark i in location x_i for input facial image I (section 3.1) computed by CEN. $\mathcal{R}$ is the regularization enforced by Point Distribution Model (Section 3.2).

3.1. Convolutional Experts Network

The first and most important step in CE-CLM algorithm is to compute a response map that helps to accurately localize

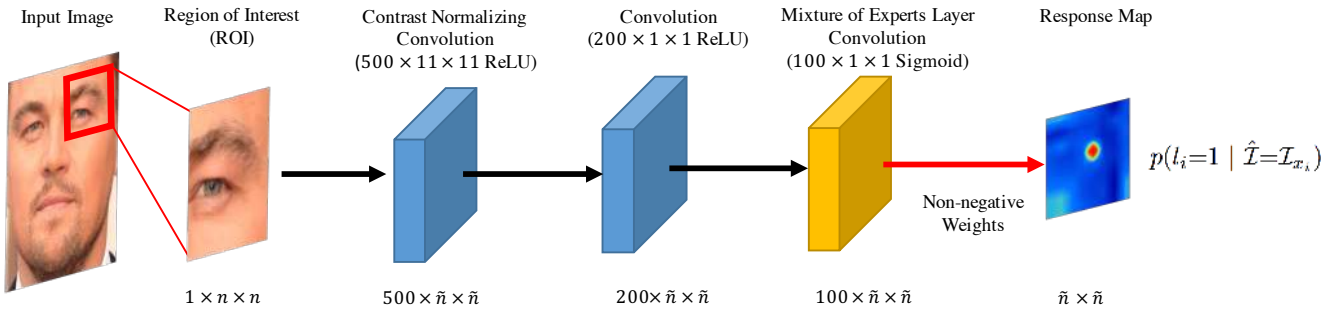


Figure 2: Overview of Convolutional Experts Network. Input image is given and based on the estimate of the landmark position a Region of Interest with size $n \times n$ is extracted from it. This small region goes through a Contrast Normalizing Convolutional layer with kernel shape $500 \times 11 \times 11$ which performs Z-score normalization before correlation operation that outputs a $500 \times \tilde{n} \times \tilde{n}$ where $\tilde{n} = n - 10$. Afterwards, the response maps are input to a convolutional layer of $200 \times 1 \times 1$ with ReLU units. Mixture of Expert Layer (ME-layer) learns an ensemble to capture ROI variations and uses a convolutional layer of $100 \times 1 \times 1$ sigmoid probability decision kernels. The output response map is a non-negative and non-linear combination of neurons in ME-layer using a sigmoid activation.

individual landmarks by evaluating the landmark alignment probability at individual pixel locations. In our model this is done by CEN which takes a $n \times n$ pixel region of interest (ROI) around the current estimate of a landmark position as input and outputs a response map evaluating landmark alignment probability at each pixel location. See Figure 2 for an illustration.

In CEN the ROI is first convolved with a contrast normalizing convolutional layer with shape $500 \times 11 \times 11$ which performs Z-score normalization before calculating correlation between input and the kernel. The output response map is then convolved with a convolutional layer of $200 \times 1 \times 1$ ReLU neurons.

The most important layer of CEN has the ability to model the final alignment probability through a mixture of experts that can model different landmark appearance prototypes. This is achieved by using a special neural layer called Mixture of Expert Layer (ME-layer) which is a convlutional layer of $100 \times 1 \times 1$ using sigmoid activation outputting individual experts vote on alignment probability (since sigmoid can be interpreted as probability). These response maps from individual experts are then combined using non-negative weights of the final layer followed by a sigmoid activation. This can be seen as a combination of experts leading to a final alignment probability. Our experiments show that ME-layer is crucial for performance of the proposed Convolutional Experts Network.

ME-layer does not include any pooling layers since pooling reduces the resolution and hurts the alignment precision. Furthermore we use convolution of size 1×1 to be able to precisely locate the landmark position as well as ensuring computational efficiency of ME-layer.

In simple terms, CEN is given an image ROI at iteration t of Equation 1 as input and outputs a probabilistic response map evaluating individual landmark alignment. Thus fitting the landmark i in position x_i follows the equation:

$$\pi^i_{x_i} = p(l_i = 1, \hat{\mathcal{I}} = \mathcal{I}_{x_i}) \tag{2}$$

l_i is an indicator for landmark number i being aligned. $\hat{\mathcal{I}}$ is the image ROI at location x_i for the image $\mathcal{I}$. The response maps π^i (of size $\tilde{n} \times \tilde{n}$) are then used for minimizing Equation 1. The detailed network training procedure is presented in Section 4.1 including chosen parameters for n at train and test time. Our experiments show that ME-layer is crucial and that making CEN model deeper does not change the performance.

3.2. Point Distribution Model and optimization

Point Distribution Models [10, 25] are used to both control the landmark locations and to regularize the shape in CE-CLM framework. Irregular shapes for final detected landmarks are penalized using the term $\mathcal{R}(\mathbf{p})$ in the Equation 1. Landmark locations $\mathbf{x}_i = [x_i, y_i]^T$ are parametrized using $\mathbf{p} = [s, \mathbf{t}, \mathbf{w}, \mathbf{q}]$ in the following 3D PDM Equation:

$$\mathbf{x}_i = s \cdot R_{2D} \cdot (\bar{\mathbf{x}}_i + \mathbf{\Phi}_i \mathbf{q}) + \mathbf{t} \tag{3}$$

where $\bar{\mathbf{x}}_i = [\bar{x}_i, \bar{y}_i, \bar{z}_i]^T$ is the mean value of the i^{th} landmark, $\mathbf{\Phi}_i$ a $3 \times m$ principal component matrix, and $\mathbf{q}$ an m-dimensional vector of non-rigid shape parameters; s, R and $\mathbf{t}$ are the rigid parameters: s is the scale, R is a 3×3 rotation matrix defined by axis angles $\mathbf{w} = [w_x, w_y, w_z]^T$ (R_{2D} are the first two rows of this matrix), and $\mathbf{t} = [t_x, t_y]^T$ is the translation.

Equation 1 can be optimized using Non-Uniform Regularized Landmark Mean Shift (NU-RLMS) [2]. Given an initial CE-CLM parameter estimate $\mathbf{p}$, NU-RLMS iteratively finds an update parameter $\Delta\mathbf{p}$ such that $\mathbf{p}^* = \mathbf{p}_0 + \Delta\mathbf{p}$,

Table 1: Comparison between CEN, LNF [2] and SVR [25] using square correlation r^2 (higher is better) and RMSE (lower is better). To evaluate the necessity of the ME-layer we also compare to CEN (no ME-layer), a model with no non-negative constraint on the weights of ME-layer. Performance drop signals the crucial role of ME-layer.

Detector	r^2	RMSE $*10^3$
SVR [25]	21.31	66.8
LNF [2]	36.57	59.2
CEN	**64.22**	**37.9**
CEN (no ME-layer)	23.81	65.11

approaches the solution of Equation 1 using Tikhonov regularized least squares.

4. Experiments

In our experiments we first evaluate the performance of Convolutional Experts Network and compare the performance with LNF [2] and SVR [25] local detectors. We also evaluate the importance of the ME-layer for CEN performance. Our facial landmark detection experiments explore the use of our model in facial landmark detection in images in especially challenging *in-the-wild* setting and with large variations in head pose, followed by experiment on 84-point landmark detection. The CE-CLM and CEN training codes are publicly available on `https://github.com/A2Zadeh/CE-CLM` and `multicomp.cs.cmu.edu/ceclm/`.

4.1. CEN Experiments

In this section we first describe training and inference methodology of the CEN local detector. We then compare the performance of CEN with LNF [2] and SVR [25] patch experts followed by an ablation study of the ME-layer.

Training Procedure: for the following experiments CEN was trained on LFPW and Helen training sets as well as Multi-PIE dataset (see Section 4.2.1). During training, we place a Normal distribution centered around the ground truth location of the response map as the label. We use 19×19 ROI with a random offset from the ground truth location. A total of 6000 such regions (meaning there were $\approx 5 \times 10^5$ alignment samples) were used during training, and we used 2000 for testing. We trained 28 sets of CENs per landmark: at seven orientations $\pm 70°, \pm 45°, \pm 20°, 0$ yaw; and four scales 17, 23, 30, and 60 pixel of interocular distance. To reduce the number of local detectors that needed to be trained we mirrored the local detectors at different yaw angles and used the same expert for left and right side of the face of the frontal view. ME-layer non-negative weights constraint was enforced using weight clipping during training. The optimizer of CEN was Adam ([16]) with small learning rate

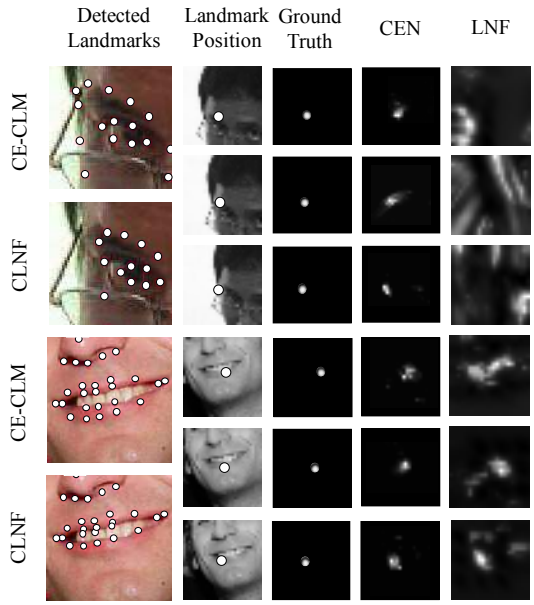


Figure 3: Comparison between response maps of CEN local detector and LNF patch experts across different landmarks. CEN shows better localization as the landmark probability is concentrated around the correct position of the landmark.

of 5×10^{-4} and trained for 100 epochs with mini-batches of 512 (roughly 800,000 updates per landmark). Training each CEN model takes 6 hours on a GeForce GTX Titan X but once trained inference can be quickly done and parallelized.

Experiments: we compare the performance of CEN local detectors over LNF and SVR patch experts. Table 1 shows the average performance for each individual landmark. Since alignment probability inference is a regression task we use squared Pearson correlation (r^2) and RMSE between the ground truth on test set and local detector output as a measure of accuracy. The train and test data for all the models are the same. On average CEN local detector performs 75.6% better than LNF and almost 200% better than SVR (calculated over r^2), which shows a significant improvement. While this is an average, for certain landmarks, views and scales performance improvement is more than 100% over LNF. This is specifically the case for 17 pixel interocular distance scale since the CEN is able to model the location of landmark based on a bigger appearance of landmark neighborhood in the image (more context present in the image).

We also evaluate the importance of the ME-layer in the CEN model in Table 1 called CEN (no ME-layer). We show that removing the non-negative constraint from the connection weights to final decision layer (essentially removing the model's capability to learn mixture of experts) and retraining the network drops the performance significantly, almost to the level of SVR. This signals that ME-layer is a crucial and possibly the most important part of CEN model capturing

ranges of variation in texture, illumination and appearance in the input support region while removing it removes the model's capability to deal with these variations.

In Figure 3 we visualize the improvement of CEN over LNF local detectors across different landmarks such as eyebrow region, lips and jaw outline. The ground truth response map is a normal distribution centered around the position of the landmark. The output response map from CEN shows better certainty about the position of the landmark as its response map is more concentrated around the ground truth position compared to LNF.

4.2. CE-CLM Experiments

In this section we first describe the datasets used to train and evaluate our CE-CLM method for facial landmark detection in images. We first describe the datasets used, followed by the baselines employed and the results of the experiments.

4.2.1 Datasets

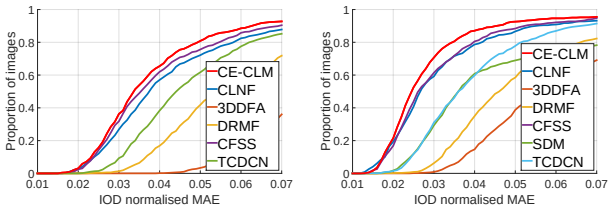
We evaluate our CE-CLM on two publicly available datasets: one within-dataset evaluation (300-W), one cross-dataset evaluation (tested on Menpo training fold) and one within (tested on Menpo test). We believe that the cross-dataset evaluation provides the strongest evidence for our model performance.

300-W [22, 24] is a meta-dataset of four different facial landmark datasets: Annotated Faces in the Wild (AFW) [43], iBUG [23], and LFPW + Helen [4, 19] datasets. We used the full iBUG dataset and the test partitions of LFPW and HELEN. This led to 135, 224, and 330 images for testing respectively. They all contain uncontrolled images of faces *in the wild*: in indoor-outdoor environments, under varying illuminations, in presence of occlusions, under different poses, and from different quality cameras. We use the LFPW and HELEN test sets together with iBUG for model evaluation (as some baselines use AFW for training).

Menpo Benchmark Challenge [34] dataset is a very recent comprehensive multi-pose dataset for landmark detection in images displaying arbitrary poses. The training set consists of 8979 images, of which 2300 are profile images labeled with 39 landmark points; the rest of the images are labeled with 68 landmarks. The test set consists of 7281 images (1946 profile), the labels for these images are not available and results are provided by challenge organizers. We only used the Menpo dataset for training our model for the challenge submission and not when comparing to the other baselines.

4.2.2 Baselines

We compared our approach to a number of established baselines for the facial landmark detection task, including both cascaded regression and local model based approaches. In all cases we use author provided implementations, meaning we compare to the best available version of each baseline and using the same methodology.



(a) With face outline (68) (b) Without face outline (49)

Figure 4: Cumulative error curves of IOD normalized facial landmark detection errors on the **300-W** test set – Helen, LFPW, and iBUG. CE-CLM performs better than all other approaches, especially in the difficult 68 landmark case that a number of cascaded regression approaches avoid. Best viewed in color.

CFSS [41] – Coarse to Fine Shape Search is a recent cascaded regression approach. It is the current state-of-the-art approach on the 300-W competition data [22, 8]. The model is trained on Helen and LFPW training sets and AFW.

CLNF is an extension of the Constrained Local Model that uses Continuous Conditional Neural Fields as patch experts [3]. The model was trained on LFPW and Helen training sets and CMU Multi-PIE [12].

PO-CR [29] – is a recent cascaded regression approach that updates the shape model parameters rather than predicting landmark locations directly in a projected-out space. The model was trained on LFPW and Helen training sets.

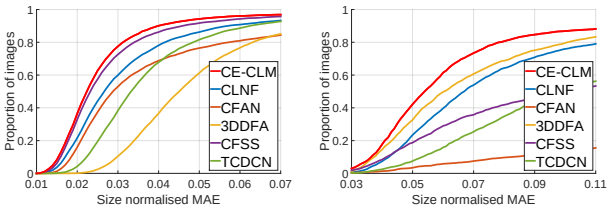
DRMF – Discriminative Response Map Fitting performs regression on patch expert response maps directly rather than using optimization over the parameter space. We use the implementation provided by the authors [1] that was trained on LFPW [4] and Multi-PIE [12] datasets.

3DDFA – 3D Dense Face Alignment [42] has shown state-of-the-art performance on facial landmark detection in profile images. The method uses the extended 300W-LP dataset [42] of synthesized large-pose face images from 300-W.

CFAN – Coarse-to-Fine Auto-encoder Network [37], uses cascaded regression on auto-encoder visual features that was trained on LFPW, HELEN and AFW.

TCDCN – Tasks-Constrained Deep Convolutional Network [40], is another deep learning approach for facial landmark detection that uses multi-task learning to improve landmark detection performance.

SDM – Supervised Descent Method is a very popular cascaded regression approach. We use implementation from the authors [30] that was trained on the Multi-PIE and LFW [14] datasets.



(a) Frontal faces (b) Profile faces

Figure 5: Results of our facial landmark detection on the **Menpo** dataset. CE-CLM outperforms all of the baselines in both the frontal and profile image case, with a very large margin in the latter. Best viewed in color.

Table 2: The IOD normalized median error of landmark detection on the **300-W** dataset. We use the typical split: Comm. – Helen and LFPW, Diff. - iBUG.

	With outline (68)		Without outline (49)	
Approach	Comm.	Diff.	Comm.	Diff.
CLNF [3]	3.47	6.37	2.51	4.93
SDM [30]	-	-	3.31	10.73
CFAN [37]	-	8.38	-	6.99
DRMF [1]	4.97	10.36	4.22	8.64
CFSS [41]	3.20	5.97	2.46	4.49
PO-CR [29]	-	-	2.67	**3.33**
TCDCN [40]	4.11	6.87	3.32	5.56
3DDFA [42]	7.27	12.31	5.17	8.34
CE-CLM	**3.15**	**5.31**	**2.30**	3.86

All of the above baselines were trained to detect either landmarks *without face outline* (49 or 51), or *with face outline* (66 or 68). For each comparison we used the biggest set of overlapping landmarks as all the approaches share the same subset of 49 feature points. For evaluating detections on profile images (present in Menpo dataset), we use the subset of shared landmarks in ground truth images and detected ones. Since the annotations of Menpo profile faces differ slightly from the 68 landmark scheme we unify them by removing the two chin landmarks and using linear interpolation to follow the annotated curve to convert the 4 eyebrow landmarks to 5; and 10 face outline landmarks to 9. This still constitutes a fair comparison as none of the approaches (including ours) were trained on Menpo. To map back to the Menpo style annotation for profile faces (for Menpo test evaluation) we use a linear least squares regressor that mapped from our 37 point annotation to the original 39 point one.

4.2.3 Experimental setup

We use the same CEN multi-view and multi-scale local detectors as described in Section 4.1. Our PDM was trained on Multi-PIE and 300-W training datasets, using non-rigid structure from motion [27]. For model fitting we use a multi-scale approach, with a higher scale CEN used for each iteration. For each iteration we use a progressively smaller Region of Interest – $\{25 \times 25, 23 \times 23, 21 \times 21, 21 \times 21\}$. For NU-RLMS we set $\sigma = 1.85, r = 32, w = 2.5$ based on grid-search on the training data. Given a bounding box, we initialized CE-CLM landmark locations at seven different orientations: frontal, $\pm 30°$ yaw, and $\pm 30°$ pitch, and $\pm 30°$ roll (we add two extra initializations $\pm 60°, \pm 90°$ yaw for Menpo dataset due to large presence of profile faces). We pick the landmarks with highest alignment probabilities as the final detection. This allows us to use a single system to perform both frontal and profile landmark detection without the need to know in advance if the face is profile.

For fairness of model comparison, the baselines and our model have been initialized using the same protocol. For 300-W dataset we initialized all of the approaches using the bounding boxes provided by the challenge organizers. For Menpo we initialized the approaches using a Multi-Task Convolutional Neural Network [38] face detector, which was able to detect faces in 96% of images. We performed an affine transformation of the bounding box to match that of bounding box around the 68 facial landmarks.

Table 3: The size normalized median landmark error on the **Menpo** dataset. We present results for profile and frontal images separately, note how our approach outperforms all of the baselines in both frontal and profile images.

	With outline (68)		Without outline(49)	
Approach	Frontal	Profile	Frontal	Profile
CLNF [3]	2.66	6.68	2.10	4.43
SDM [30]	-	-	2.54	36.73
CFAN [37]	2.87	25.33	2.34	28.09
DRMF [1]	-	-	3.44	36.14
CFSS [41]	2.32	9.99	1.90	8.42
PO-CR [29]	-	-	2.03	36.04
TCDCN [40]	3.32	9.82	2.81	8.69
3DDFA [42]	4.51	6.02	3.59	5.47
CE-CLM	**2.23**	**5.39**	**1.74**	**3.32**

4.2.4 Landmark Detection Results

As common in such work, we use commutative error curves of size normalized error per image to display landmark detection accuracy. We also report the size normalized median per image error. We report the median instead of the mean as the errors are not normally distributed and the mean is very susceptible to outliers. For 300-W dataset we normalize the error by inter-ocular distance (IOD), for Menpo dataset (containing profile faces) where one of the eyes might not be visible we instead use the average of width and height of the

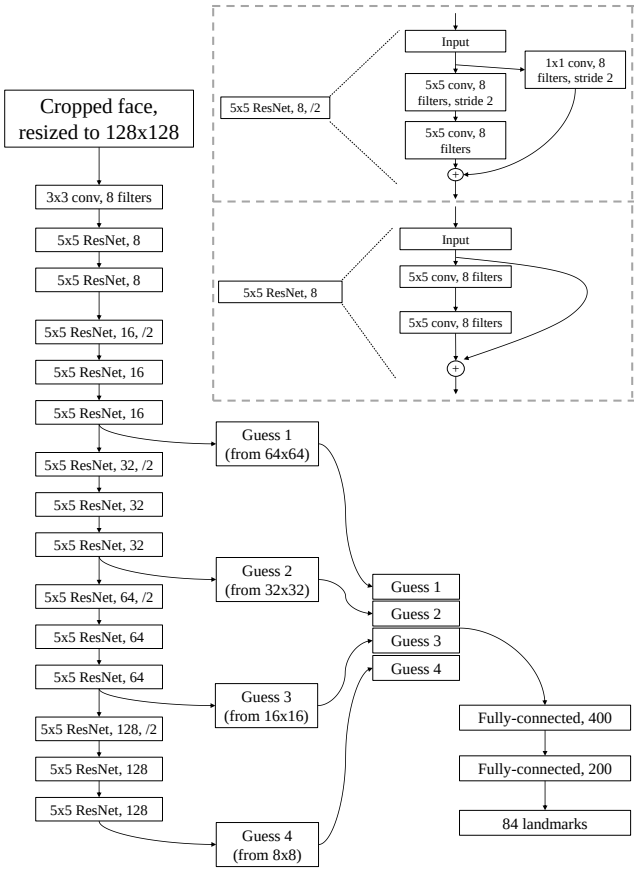


Figure 6: The architectuere of the Correction Network.

face on training set and diagonal size of the face on test set.

Results of landmark detection on the **300-W** dataset can be seen in Table 2 and Figure 4. Our approach outperforms all of the baselines in both the 68 and 49 point scenarios (except for PO-CR in the 49 landmark case on the iBUG dataset). The improved accuracy of CE-CLM is especially apparent in the 68 landmark case which includes the face outline. This is a more difficult setup due to the ambiguity of face outline and which a lot of approaches (especially cascade regression based ones) do not tackle.

We perform a cross-dataset experiment were we train the model without any training data from Menpo2D [36]. We evaluate the model on Menpo train set and compare to a number of state-of-the-art approaches. Figure 5 and Table 3 show our model outperforming all of the baselines on this difficult task. The performance improvement is especially large on profile faces, which SDM, CFAN, DRMF, and PO-CR approaches are completely unable to handle. We also outperform the very recent 3DDFA model which was designed for large pose face fitting. As these results are on a cross-dataset evaluation, they demonstrate how well our method generalizes to unseen data and how well it performs

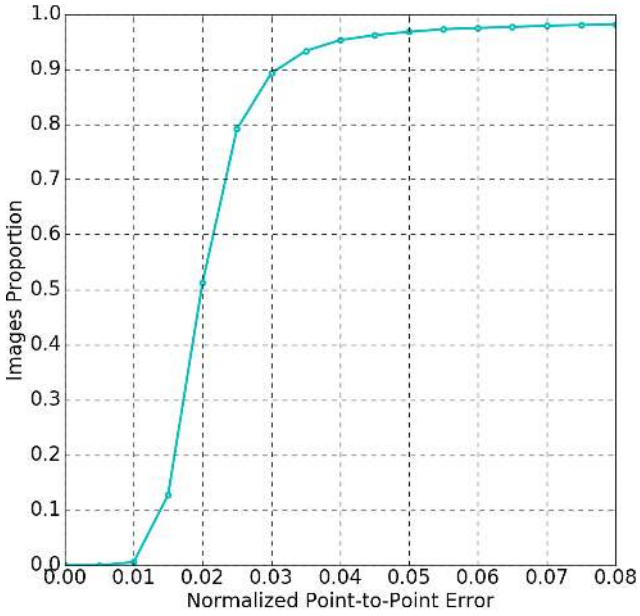


Figure 7: Results of our 84-points dense facial landmark detection on the withheld **Menpo3D** test dataset.

on challenging profile faces.

4.3. 84-points Landmark Detection Experiments

Since CE-CLM extracts 3D landmark position by nature, we use it as the cornerstone of our approach. We use two key networks to adapt the 68 3D outputs of CE-CLM to the 3D annotations of the Menpo3D dataset. The first network is called the Adjustment Network which is a fully connected deep neural network with ReLU activations. The second network is a Deep Residual network named Correction Network. The Correction Network learns dataset specific priors and correction for CE-CLM since CE-CLM is not trained on Menpo or Menpo3D.

Adjustment Network: is a fully connected deep neural network with 3 layers. The first and second layer are 200 ReLU units. The third layer is the final prediction layer with ReLU activation which regresses the landmark positions with 84×3 units; (x, y, z) for each landmark position. The input to the adjustment network is the concatenation of output of CE-CLM with dimension of (68×3) and output of Correction Network with dimension of (84×3) units.

Correction Network: We use a deep residual network [13] for the Correction Network. The input to the Correction Network is a 128×128 reshaped face box. We use MTCNN face detector [39] to extract the faces in the train and test sets. The architecture of the Correction Network is shown in Figure 6 with ReLU activations. The output of the Correction

Figure 8: Examples from CE-CLM output on Menpo3D dataset.

Network is connected to the adjustment network.

As the first step, we performed landmark detection using CE-CLM and used the detection results as input to the Adjustment Network. Both the Adjustment and Correction network were trained together using Adam [17] with learning rate of 5×10^{-4}. Both networks were trained using all the Menpo3D trainset, except for the data points randomly chosen for validation set. Batch normalization was used after every convolutional layer in residual networks. Both networks were trained for 125 epochs. We also retrain the CEN network using Menpo train data. We observed that simply using Menpo train data and starting from randomly initialized CEN network weights does not achieve satisfactory results. However, using Curriculum Learning paradigm [9], by starting from 300-W and Multi-PIE which are easier datasets and then moving to Menpo training data after 100 epochs, we were able to successfully train the CEN network for another 50 epochs. The results of our approach on Menpo3D test data is reported in Figure 7. For example fits see Figure 8.

5. Conclusion

In this paper we used the CE-CLM approach for facial landmark detection in Menpo3D dense 84-points landmark detection. We reiterated the state-of-the-art performance of CE-CLM approach using multiple experiments. For Menpo3D challenge, since the CE-CLM approach outputs 3D landmark positions by default, we used a network, called Adjustment Network, to map the 68 markup to 84 markup (proposed by Menpo3D challenge). We furthermore used a Deep Residual Network, called Correction Network, to learn dataset specific corrections for Menpo3D data since CE-CLM is not trained on Menpo3D. The results of our experiment on Menpo3D test set is presented in figure 7.

Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA 2014-14071600011. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon.

References

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

[2] T. Baltrusaitis, L.-P. Morency, and P. Robinson. Constrained local neural fields for robust facial landmark detection in the

wild. In *IEEE International Conference on Computer Vision Workshops*, 2013.

[3] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Continuous conditional neural fields for structured regression. In *Computer Vision–ECCV 2014*, pages 593–608. Springer, 2014.

[4] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011.

[5] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. *SIGGRAPH*, 1999.

[6] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou. 3d face morphable models" in-the-wild". *arXiv preprint arXiv:1701.05360*, 2017.

[7] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by Explicit Shape Regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2887–2894. Ieee, jun 2012.

[8] G. G. Chrysos, E. Antonakos, P. Snape, A. Asthana, and S. Zafeiriou. A Comprehensive Performance Evaluation of Deformable Face Tracking "In-the-Wild". 2016.

[9] V. Cirik, E. Hovy, and L.-P. Morency. Visualizing and understanding curriculum learning for long short-term memory networks. *arXiv preprint arXiv:1611.06204*, 2016.

[10] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *TPAMI*, 23(6):681–685, Jun 2001.

[11] D. Cristinacce and T. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, 2006.

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

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. 2007.

[15] K. KangGeon, T. Baltrušaitis, A. Zadeh, L.-P. Morency, and G. Medioni. Holistically constrained local model: Going beyond frontal poses for facial landmark detection. In *British Machine Vision Conference (BMVC)*, 2013.

[16] D. Kingma and J. Ba. Adam: A method for stochastic optimization https://arxiv.org/abs/1412.6980.

[17] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 365–372, 2009.

[19] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *Computer Vision–ECCV 2012*, pages 679–692. Springer, 2012.

[20] B. Martinez and M. Valstar. Advances, challenges, and opportunities in automatic facial expression recognition. In B. S. M. Kawulok, E. Celebi, editor, *Advances in Face Detection and Facial Image Analysis*, pages 63 – 100. Springer, 2016.

[21] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency. Context-dependent sentiment analysis in user-generated videos. In *Association for Computational Linguistics*, 2017.

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

[23] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV*, 2013.

[24] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W), Workshop on Analysis and Modeling of Faces and Gestures*, 2013.

[25] J. Saragih, S. Lucey, and J. Cohn. Deformable Model Fitting by Regularized Landmark Mean-Shift. *IJCV*, 2011.

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

[27] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *TPAMI*, 30(5):878 –892, may 2008.

[28] 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 *CVPR*, 2016.

[29] G. Tzimiropoulos. Project-Out Cascaded Regression with an application to Face Alignment. In *CVPR*, 2015.

[30] X. Xiong and F. Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[31] A. Zadeh, T. Baltrušaitis, and L.-P. Morency. Convolutional experts network for facial landmark detection.

[32] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. Tensor fusion network for multimodal sentiment analysis. In *Empirical Methods in NLP*, 2017.

[33] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *2016 IEEE Intelligent Systems*, 2016.

[34] S. Zafeiriou. The menpo facial landmark localisation challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[35] S. Zafeiriou, G. Chrysos, G. Trigeorgis, and J. Deng. The 3D Menpo Facial Landmark Tracking Challenge. *ICCVW*, 2017.

[36] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen. The Menpo Facial Landmark Localisation Challenge: A step closer to the solution. *CVPRW*, 2017.

[37] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *ECCV*. Springer, 2014.

[38] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.

[39] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[40] Z. Zhang, P. Luo, C.-C. Loy, and X. Tang. Facial Landmark Detection by Deep Multi-task Learning. In *ECCV*, 2014.

[41] S. Zhu, C. Li, C. C. Loy, and X. Tang. Face Alignment by Coarse-to-Fine Shape Searching. In *CVPR*, 2015.

[42] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face Alignment Across Large Poses: A 3D Solution. In *CVPR*, 2016.

[43] 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.