Unconstrained Face Alignment via Cascaded Compositional Learning

Shizhan Zhu[1,2] Cheng Li[2] Chen Change Loy[1,3] Xiaoou Tang[1,3]
[1]Department of Information Engineering, The Chinese University of Hong Kong
[2]SenseTime Group Limited
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
zs014@ie.cuhk.edu.hk, chengli@sensetime.com, ccloy@ie.cuhk.edu.hk, xtang@ie.cuhk.edu.hk

Abstract

We present a practical approach to address the problem of unconstrained face alignment for a single image. In our unconstrained problem, we need to deal with large shape and appearance variations under extreme head poses and rich shape deformation. To equip cascaded regressors with the capability to handle global shape variation and irregular appearance-shape relation in the unconstrained scenario, we partition the optimisation space into multiple domains of homogeneous descent, and predict a shape as a composition of estimations from multiple domain-specific regressors. With a specially formulated learning objective and a novel tree splitting function, our approach is capable of estimating a robust and meaningful composition. In addition to achieving state-of-the-art accuracy over existing approaches, our framework is also an efficient solution (350 FPS), thanks to the on-the-fly domain exclusion mechanism and the capability of leveraging the fast pixel feature.

1. Introduction

Face alignment [8, 1, 7, 28, 29, 31, 42, 25, 35, 21] aims to automatically localise facial parts locations, which are essential for many subsequent processing modules, such as face recognition [24], face attributes prediction [10], and robust face frontalisation [16]. The objective of this study is to devise an effective and efficient face alignment method to handle faces with unconstrained variations. AFLW dataset [20] provides a good example of images typically found in unconstrained scenarios. The dataset is extremely challenging given the shape and appearance variations – it contains in-the-wild faces obtained from Flickr, with rich face expressions, and head poses up to ±120° for yaw and ±90° for pitch and roll. Some examples of faces are shown in Fig. 1.

The study of face alignment has made rapid progresses in recent years. But how effective can existing methods handle unconstrained faces? To gain a preliminary insight, we plot the error distribution of existing methods on the full AFLW dataset. We first select the Supervised Descent Method (SDM) [36], a representative method among the mainstream cascaded regression approaches [11, 5, 36, 27, 33]. As shown in Fig. 1(a), even the approach is retrained on AFLW, its effective scope is confined within frontally biased faces, and it has difficulty to cover an enlarged shape parameter space due to large head rotations and face deformations caused by rich expressions. Xiong and De la Torre [37] have the same observation – a cascaded regressor such as Supervised Descent Method (SDM) is only effective within a specific domain of homogeneous descent (DHD)[1].

We make a second attempt with an intuitive multi-view approach – first estimating head poses then followed by face alignment on a specific view[2]. The performance improves but as revealed in Fig. 1(b), the heuristic partitioning with respect to only the head pose is still suboptimal because it neglects other shape deformation or appearance variations, e.g. large mouth, large face scale or sunglasses. Moreover, it assumes independence between different view models without considering their inter-complementary and regularisation role. Hence, the error caused by head pose estimation could easily be propagated and amplified to the final shape estimation, reducing the overall robustness.

The preliminary tests show the difficulties of covering a wider range of shape and appearance variations beyond frontal faces, either with a single or multiple models. A few recent studies [37, 34, 18, 17, 32] have started to work on this important and relatively unexplored problem of face alignment. These studies mainly resort to three-dimensional (3D) face modelling [18] or constrain the problem with additional assumptions, e.g., adding temporal prior [37] or

[1] A DHD refers to optimisation spaces of a function that share similar directions of gradients.

[2] We adopt a state-of-the-art head pose estimator [15]. Given a test face image, we use the estimated head pose to select the best-match from a set of view-specific SDM cascaded regressors and the associated initial shape for alignment.

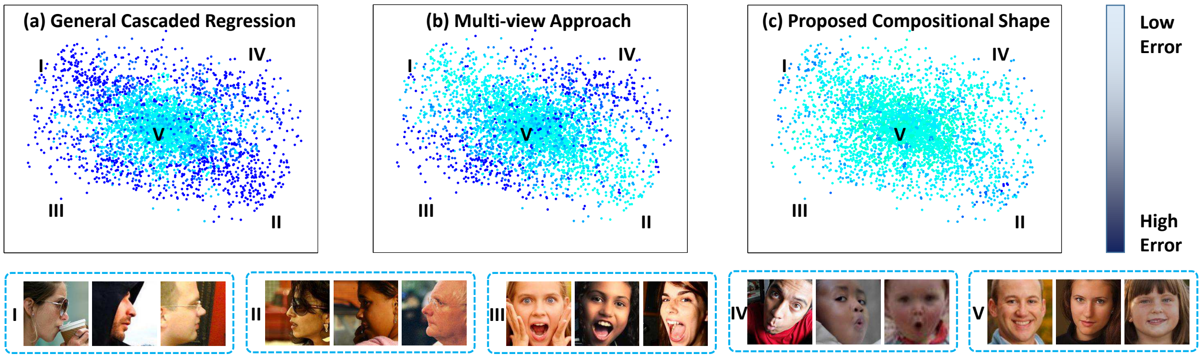


Figure 1. Test error distributions of three approaches on the AFLW dataset [20]. We select two factors, yaw and mouth size, to visualise the distribution and provide the representative facial images in five regions (I-V). It is observed that a cascaded regression approach only obtains low errors within the region of frontal faces (V). A multi-view approach does not fully address the global shape and appearance variations. Our cascaded compositional learning generates satisfying results among all the faces.

confine the problem scope to only frontal and absolute profile views [34] (see Sec. 2 for discussion).

In this paper, we propose an effective and efficient alternative for unconstrained face alignment. It does not rely on 3D face modelling and 3D annotations, and does not make assumption on the pose range. It can comfortably deal with arbitrary view pose and rich expressions in the full AFLW dataset. In addition, the alignment is achieved on a single image without the need of temporal prior. We show that all these appealing properties can be achieved through reformulating the popular cascaded regression scheme into a cascaded compositional learning (CCL) problem. Specifically, instead of using a single cascaded regressor to cover the global shape parameter space, we follow [37] to partition the optimisation space into multiple domains of homogeneous descent (DHDs). Each domain is handled by a domain-specific cascaded regressor. Given a test image, its shape is collectively estimated as a compositional shape from various DHDs. We highlight some unique properties of CCL:

(i) *Robust compositional prediction*: we estimate a compositional vector in a decision trees framework with a novel splitting function. The function is formulated such that it optimises directly the landmark locations, thus leading to accurate alignment given arbitrary poses and face deformations (Fig. 1(c)). We note that this is not a naive mixture estimation – we need to ensure the trees to estimate a meaningful composition that captures unique correlations between domains so that the estimated compositional vector allows semantically nearer domains to enjoy similar recommendation than those further away in the domain space.

(ii) *Fast speed*: We formulate our approach to discard unpromising domain(s) on-the-fly. In addition, the approach is designed to leverage the fast pixel difference feature [5]. Our method achieves 350 FPS on a single core desktop, which is comparable to existing face alignment methods, yet 10 times faster than the heuristic multi-view strategy based on a recent fast head pose estimation [23].

Extensive empirical results indicate that our method outperforms existing methods on challenging datasets with large shape and appearance variations, e.g., AFLW [20] and AFW [45]. The codes and the supplementary annotations on AFLW [20] are available at mmlab.ie.cuhk.edu.hk/projects/compositional.html.

2. Related Work

Unconstrained face alignment beyond frontally biased faces is an emerging research topic [37, 34, 18, 17, 32]. Jourabloo et al. [18] propose a 3D approach that employs cascaded regression to predict coefficients of 3D base shapes and a 3D-2D projection matrix. We show in the experiments that optimising the base shape coefficients and projection is indirect and sub-optimal since smaller parameter errors are not necessarily equivalent to smaller alignment errors [5]. Tulyakov et al. [32] extend the regression target to the 3D space. Nevertheless, this method relies on 3D databases with limited shape variations, and cannot utilise existing 2D in-the-wild databases, e.g. AFLW [20] with 20K+ faces. Wu et al. [34] propose an occlusion-robust cascaded regressor to handle extreme head poses and occlusion. Their method, however, is only verified on less challenging images with lab environment, and it lacks specific mechanism to handle arbitrary shape variation other than two standard profile views. Hsu et al. [17] extended the mixture of tree model [45] to achieve better accuracy and efficiency. However it assumes face shape to be a tree structure, enforcing strong constraints toward shape variation. Xiong et al. [37] pointed out that standard cascaded regression approaches such as the SDM [36] tend to average conflicting gradient directions resulting in undesirable face alignment performance. To overcome the weakness of lo-

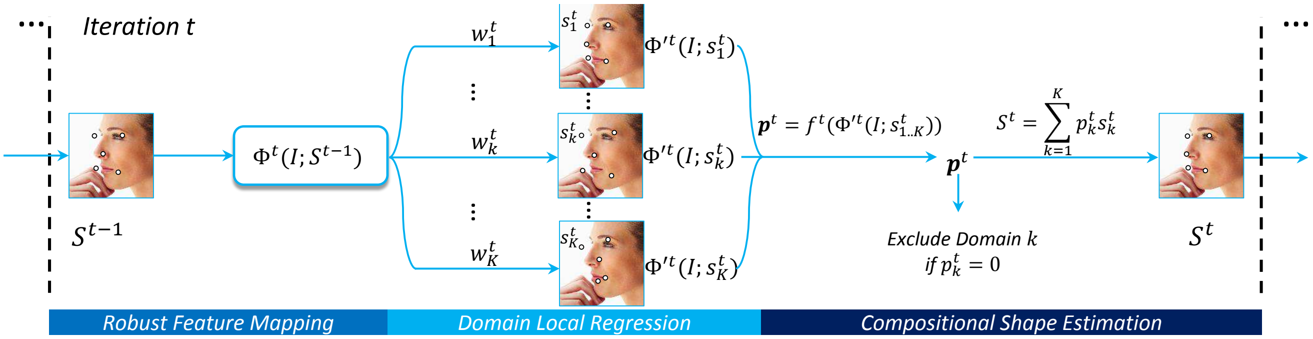


Figure 2. Illustration for one testing iteration Cascaded Compositional Learning (CCL). It contains three steps: 1) Robust feature mapping to obtain shape indexed feature $\Phi^t(I; S^{t-1})$; 2) Each domain k obtains regressed shape s_k^t and the corresponding feature $\Phi'^t(I; s_k^t)$; 3) Predicting the composition $\mathbf{p}^t$ to obtain the estimated shape S^t. Accordingly, our model consists of three components for each iteration: feature extraction Φ^t, a set of local regressors $W^t = \{w_1^t, ..., w_k^t, ..., w_K^t\}$, and the composition estimator f^t. With the on-the-fly domain exclusion mechanism, in Step 2, we bypass the computation of s_k^t and $\Phi^t(I; s_k^t)$ if the k-th domain has been excluded before the iteration. The value of s_k^t and $\Phi^t(I; s_k^t)$ are kept fixed after the exclusion of the k-th domain.

cal cascaded regression methods, they partition the optimisation space into several DHDs and learn separate descent maps for each domain. However, the decision of picking the suitable domain strongly depends on the estimated shape of previous frame in the video. When applied to the static images, it lacks a principled method to accurately locate the correct domain. Our method addresses this limitation and provides a solution to select (and exclude) domains through cascaded compositional learning.

3. Cascaded Compositional Learning

Given a face image I, a face alignment algorithm aims to predict the facial shape S, i.e., the x, y coordinates $[x_1, y_1, ..., x_l, y_l, ..., x_L, y_L]$ for the L facial landmarks. The cascaded regression approach (e.g. [36, 27]) consists of T iterations, each of which updates the shape via

$$S^t = S^{t-1} + W^t \Phi^t(I; S^{t-1}), \tag{1}$$

where a learned linear regressor W^t is used to map the shape indexed feature $\Phi^t(I; S^{t-1})$ to the shape residual.

To address the limitation of local regression when applied to a problem with large global shape variations, we propose to represent our shape prediction in each iteration as a composition of shape estimations across multiple DHDs [37]. More precisely, we define a set of domain-specific 'locally-descent' regressors $W^t = \{w_1^t, ..., w_k^t, ..., w_K^t\}$ and a composition estimator f^t. We predict the shape in one cascade by

$$S^t = \sum\nolimits_{k=1}^{K} p_k^t s_k^t, \quad s_k^t = S^{t-1} + \Delta s_k^t, \tag{2}$$

$$\Delta s_k^t = w_k^t \Phi^t(I; S^{t-1}), \quad \mathbf{p}^t = f^t(\Phi'^t(I; s_{k=1...K}^t)), \tag{3}$$

where both the descent of each domain Δs_k^t and the composition vector $\mathbf{p}^t = [p_1^t, ..., p_k^t, ..., p_K^t]^\top$ are based on our estimation from the shape indexed feature (Eq. 3 is detailed in Sec. 3.1). Here, the composition $\mathbf{p}^t$ is a meaningful quantitative description of domains. For example, the composition of two incompatible domains (e.g. left and right profile-view domains) should not co-occur. Each composition is also non-negative that provides valid shape contribution. We point out that the composition $\mathbf{p}^t$ is estimated after Δs_k^t so that it could directly exploit the local appearance $\Phi'^t(I; s_k^t)$[3]. This provides us the opportunity to handle faces in the unconstrained scenario by still only extracting the fast pixel feature throughout our approach.

Domain partition. We define the K domains by partitioning all training samples into K subsets. Following [37], we partition all samples according to the principle components of shape and local appearance. Each component halves the samples and hence K is always a power of 2. It is worth pointing out that head pose is not the only underlying factor for the partition. By observing the mean face of each domain, we found that some domains are dominant by shape deformation or appearance property, e.g. wide-open mouth, large facial scaling, large face contour or faces with sunglasses. All domains share the same feature mapping (Φ^t).

Inference. Figure 2 illustrates one iteration in the test phase of CCL. We first obtain the shape indexed feature based on the shape estimation of previous iteration. This is followed by feeding the features to all domains to obtain their estimated shape residual and the corresponding feature indexed at the updated locations. We then predict the composition of various domains by jointly considering the local appearance from all domains. We repeat the above procedure (Eq. 2, 3) for T iterations to obtain our final estimate. Note that the first iteration has no S^{t-1}, and hence, each domain begins

[3] We will show in Sec. 3.2 that the feature Φ' is a simple variant of Φ to better serve the inference of the composition.

from its domain-specific mean shape. To further accelerate the speed, we adaptively exclude incompatible domain on-the-fly. As shown in Fig. 2, if a domain k receives zero-composition estimate (i.e. $p_k^t = 0$), we instantly exclude this domain and by-pass all its computation thereafter. We empirically find that this strategy accelerates the inference procedure four times without any loss of accuracy.

Learning. We learn the feature mapping (Φ^t) and compositional shape estimation ($W^t, \mathbf{p}^t$) consecutively. Their motivation and learning steps are detailed in Sec. 3.1 and 3.2. To discuss the core step in our framework, we begin by introducing the learning of compositional shape ($W^t, \mathbf{p}^t$).

3.1. Compositional Shape Estimation

The learning of compositional shape consists of learning W^t and $\mathbf{p}^t$. Learning of W^t is straightforward. As we have assumed that each domain k exhibits homogeneous descent, the regressors w_k^t could be learned via ridge regression

$$\min_{w_k^t} \sum_{i \in \mathcal{T}_k} \|\hat{S}_i - S_i^{t-1} - w_k^t \Phi^t(I; S_i^{t-1})\|_2^2 + \gamma \|w_k^t\|_F^2. \quad (4)$$

$\mathcal{T}_k$ denotes training sample indices in the k-th domain.

The difficulties of shape prediction arise from the accurate estimation of the composition $\mathbf{p}^t$ (f^t in Eq. 3). A straightforward modeling of f^t can be a general multi-class classifier (e.g. SVC or classification forest etc.) by treating each domain as a class. In the test phase, we simply set $\mathbf{p}^t$ to be the class posterior probability estimation. However, we found that these general classification approaches generate poor results. The core reason is that they regard different domains equally. In particular, during training, they only favor a single optimal domain for each sample and cast the same loss punishment for any other domain prediction. This brings two disadvantages that lead to the poor results. First, with the existence of the *Basin of Attraction* phenomenon [37], some sub-optimal domains also provide relatively accurate s_k^t. They provide similar landmark location and local appearance as the optimal domain, but treated as negative class, confusing the training of a classifier. Second, the classification training does not heavily punish on incompatible domains (e.g. right-profile-view domain for a left-profile face). In the test phase, an inclusion of incompatible domain could seriously hurts the overall robustness.

We overcome the aforementioned problem through learning to predict a composition $\mathbf{p}$ by directly optimising the discrepancy between the compositional shape S_i and the ground-truth shape $\hat{S}_i$ (we omit the index t throughout this section for brevity)

$$\begin{aligned} &\min_f \textstyle\sum_{i \in \mathcal{T}} \|\hat{S}_i - S_i\|_2^2 \\ s.t. \quad & S_i = \textstyle\sum_{k=1}^K p_{i,k} s_{i,k}, \\ & \mathbf{p}_i = f(\Phi_i'), \mathbf{p}_i \geq 0, \|\mathbf{p}_i\|_1 = 1, \end{aligned} \quad (5)$$

where $s_{i,k}$ denotes the regressed shape of domain k for training sample i based on the learned regressor, and Φ_i' denotes all the feature $\Phi'(I; s_{i,k})(k = 1, ..., K)$. The direct optimisation toward shape enables us to readjust the punishment for different composition configuration and hence avoid the limitation of classifier learning. The enforced probabilistic simplex constraint for $\mathbf{p}$ follows the same spanning space of classification confidence estimation.

To predict a meaningful and robust composition at fast speed, we choose the forest model as the specific form of f in Eq. 5. Rather than using the classification entropy loss to direct the split learning, we propose a new splitting mechanism to enable a direct optimisation for shape over the composition $\mathbf{p}^t$ for learning the trees (Eq. 5). More precisely, we learn to split a node by jointly optimising the split parameters Θ (selected feature index and threshold) and the composition $\mathbf{p}^{(\alpha)}, \alpha = \{L, R\}$ with the following loss

$$\begin{aligned} &\min_{\Theta; \mathbf{p}^{(\alpha)}} \sum_{\alpha = L,R} \sum_{i \in \mathcal{Q}} \|\hat{S}_i - S_i\|_2^2 \\ s.t. \quad & S_i = \textstyle\sum_{k=1}^K (\mathbf{p}^{(\alpha)})_k s_{i,k}, \quad \mathbf{p}^{(\alpha)} \geq 0, \|\mathbf{p}^{(\alpha)}\|_1 = 1. \end{aligned} \quad (6)$$

$\mathcal{Q}$ denotes training sample indices in the current split node.

After learning, our tree structure includes the following learned parameters: 1) Each split or leaf node j carries a unique composition vector $\mathbf{p}_j$; 2) Each split node j carries the split parameter Θ_j; 3) Each leaf node j carries the training sample indices $\mathcal{Q}_j$ that reached that leaf during training.

In the test phase, we traverse each sample in each tree to reach one leaf node. We denote $\mathcal{Q}^*$ as the union set of all training sample indices $\mathcal{Q}_j$ in all reached leafs j. We also denote $\mathcal{K}$ as domain indices that have not been excluded before the current iteration. We obtain the composition prediction by optimising

$$\begin{aligned} &\min_{\mathbf{p}} \sum_{i \in \mathcal{Q}^*} \|\hat{S}_i - \sum_{k \in \mathcal{K}} p_k s_{i,k}\|_2^2 \\ &s.t. \quad \mathbf{p} \geq 0, \|\mathbf{p}\|_1 = 1, p_k = 0|_{k \notin \mathcal{K}}. \end{aligned} \quad (7)$$

Note that i is still the index of training samples. The optimisation can be efficiently solved by quadratic programming and we empirically found that the time cost to obtain the composition $\mathbf{p}$ is negligible compared to the feature extraction process. Having obtained the composition vector, we obtain our shape estimate by aggregating all shape composition via Eq. 2.

3.2. Learning a Suitable Feature Mapping

Φ^t: **Enrichment of feature encoding**. Feature mapping Φ^t plays important roles as it is used both for regression and composition prediction. We construct our feature representation based on the Local Binary Feature (LBF) [27] framework for its fast speed and powerful local representation learning.

The original LBF learns the local representation via growing a regression forest [3] for each landmark. We empirically found that such encoding (codebook of 2D regression offset) is insufficient when extending our problem to the unconstrained scenario, because LBF is not robust to self-occlusion, and also frequently leads to incorrect composition estimation when features are indexed at background area. Our framework requires a more proactive learning of background representation and additionally encoding of landmark visibility[4].

Hence, the negative training samples are additionally included as shown in Fig. 3(b) similar to the face detection application [6]. We augment the decision space, 2D regression offset $\pi_l \circ \Delta\hat{S}_i^t$[5], with an additional binary classification label $\hat{c}_i(l)$, describing the visibility of the landmark in the local region.[6] Similar to [12, 6], we learn to minimise the structured target loss with the Hough Forest strategy. Specifically, the probability of a node to be a regression split equals to the proportion of the positive samples in the current split node. Figure 3 provides an illustration to the enriched feature learning.

We point out two major advantages for the new feature. First, indicators in the feature better encodes the visibility information. This is a strong cue for predicting the composition $\mathbf{p}^t$ (Eq. 3). Second, this feature benefits regression accuracy with much better robustness to self-occlusion than the original LBF. It avoids heuristic occlusion pattern partition [41] or occluded feature down-weighting [34].

Φ'^t: **Learning suitable feature for composition prediction**. To further tailor the feature for composition prediction ($\mathbf{p}$ in Eq. 3), we process the obtained feature Φ by smoothing and compressing it into a $K \times L$-dimensional intermediate representation Φ' for learning the trees. This is based on our empirical finding that Φ is high-dimensional and each indicator is quite noisy. Such representation is not favorable for directing binary tests in the trees. Hence, we discard the 2D regression offset information and purifies the local visibility classification information in the feature. More precisely, for each feature $\phi_l(I; s_k)$ extracted from landmark l indexed at s_k, we learn a linear mapping to map $\phi_l(I; s_k)$ to its local visibility classification label $\hat{c}(l)$. Our experimental results show that such representation is discriminative enough for the trees to learn for composition prediction.

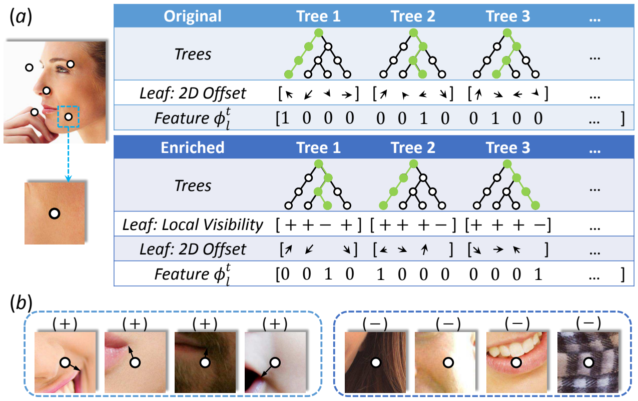


Figure 3. (a) The left mouth corner of the test face is invisible within the current local region. Trees in our feature mapping approach satisfactorily traverses the sample to the 'invisible'(−) leaf node in most of the trees, while the original LBF feature [27] contains strong noise. (b) Typical training samples used for training. The original LBF feature only considers positive case (+).

4. Experiment

Datasets. We select two types of datasets (see Tab. 1) with their most recent and challenging benchmarks. It is worth noting that we obtain competitive performance to [33, 44, 22] on 300W [28] and did not include its result due to space limitation and its saturated performance [33].

In-the-wild datasets. We select AFLW [20] and AFW [45] as the main test bed due to their challenging shape variations and significant view changes. **AFLW** contains 24386 in-the-wild faces (obtained from Flickr) with head pose up to ±120° for yaw and ±90° for pitch and roll with extremely challenging shape variations and deformations (Fig. 4(a)). Besides, AFLW also demonstrates strong external-object occlusion. There are a total of 20.6% invisible landmarks caused by self or external occlusion, larger than 13.3% on COFW [4] where only external-object-occlusion is exhibited. **AFW** is a popular benchmark and contains 468 in-the-wild up-right faces (obtained from Flickr) with yaw degree up to ±90° and also with rich shape variation (Fig. 4(b)).

AFLW provides at most 21 points for each face, but excluding coordinates for invisible landmarks, causing difficulties for training most of the existing baseline approaches. To make fair comparisons, we manually annotate the coordinates of these invisible landmarks to enable training of those baseline approaches. Our annotation does not include two ear points because it is very difficult to decide the location of invisible ears. This causes the point number of

[4] Here the 'visibility' only refers to the local region of current estimated location rather than the whole face. For example, even if the left mouth corner in Fig. 3(a) is not occluded, it is invisible within the local region given the current estimated location.

[5] Following [27], we denote the ground-truth shape for each training sample i as $\hat{S}_i$, and the target shape residual as $\Delta\hat{S}_i^t = \hat{S}_i - S_i^{t-1}$. $\pi_l \circ$ denotes the operation extracting the two elements $(2l-1, 2l)$ from the shape.

[6] We label $\hat{c}_i(l)$ to be negative for two cases: 1) the ground-truth landmark itself is invisible (e.g. occluded); 2) $\|\pi_l \circ \Delta\hat{S}_i^t\|_2$ is larger than pixel extracting radius (as the left mouth corner landmark in Fig. 3(a)). Definition of radius is similar to [27] that defines the pixel difference feature extraction scope.

Table 1. Detailed evaluation settings for our experiments.

Category	Evaluation Name	Training Set	# of Training Samples	Testing Set	# of Testing Samples	Point	Normalising Factor	Setting
In-the-wild	AFLW-PIFA	AFLW	3901	AFLW	1299	21	Face Size	Following [18]
	AFLW-Full	AFLW	20000	AFLW	4386	19	Face Size	
	AFLW-Frontal	AFLW	20000	AFLW	1165	19	Face Size	
	AFW	AFLW	3901	AFW	468	6	Face Size	Following [18]
Lab-Environment	MultiPIE	MultiPIE	1600	MultiPIE	1500	39/68	Face Size	Following [17, 45]
	FERET	MultiPIE	1600	FERET	221/244	11	Half Eye-Mouth Distance	Following [34]

AFLW-Full and AFLW-Frontal to be 19 in Tab. 1 (excluding the two ear points). Similar ear points exclusion can also be found in [39] for AFLW and [2] for LFPW. Our complementary annotation only aims to train various baseline approaches.

The original AFLW does not provide train-test partition. We thus adopt the following three settings: (I) *AFLW-PIFA*: We strictly follow train-test partition provided by PIFA [18]; (II) *AFLW-Full*: We benchmark the full set AFLW with our own partition. We randomly partition it into a disjoint 20000-image training set and 4386-image test set. Each image is annotated with *full* 19 landmarks (without two ears as aforementioned) and visibility binary label (following original source).[7] (III) *AFLW-Frontal*: We only use the frontal subset (all landmarks are visible, totally 1165 images) out of the 4386-image test set to validate whether the approach degenerates on the frontal faces.

Lab-environment datasets. To follow existing approaches [34, 17, 40, 45] that evaluate on lab-environment datasets, we choose to experiment on MultiPIE [14] and FERET [26]. **MultiPIE** was originally collected in a constrained lab environment with a setting of multiple poses, illuminance and facial expressions. A subset of 6152 faces are defined [14] and labelled with either 39 (profile-view) or 68 (frontal-view) facial landmarks. Zhu et al. [45] further annotate another 400 profile-view faces with 39 landmarks. **FERET**. originally collects 14126 images with the purpose of evaluating face recognition task. Wu et al. [34] provides 11-point annotation on a subset of 465 profile faces.

Evaluation metric. Following most previous works, we obtain the error for each test sample via averaging normalised errors for all annotated landmarks. All evaluation settings for different experiments are noted in Tab. 1. We demonstrate our results with mean error over all samples, or via Cumulative Error Distribution (CED) curve, to better compare with existing performance. We evaluate the speed of the algorithms on a single core i7-4790 CPU.

4.1. Comparison with multi-view approaches

It is necessary to conduct comparison with combination of head pose and cascaded regression. To prepare strong baselines, we choose three schemes to estimate head poses: (1) forest approach [15][8]; (2) deep model [38]; (3) correct head pose given by annotation (hence it enjoys extra advantage). We point out that both (1) and (2) are re-trained on AFLW training set because their original training set in the literature exhibits limited appearance variation or view range and receive poorer results than their re-trained models. We choose LBF [27] and CFSS [44], two state-of-the-art cascaded regression approaches, for each multi-view baseline. We conduct the comparison with the AFLW-Full evaluation (Tab. 1) because it provides the full shape variations. For baseline approach, we train 15 separate view-specific models by dividing the view space into 5 yaws and 3 pitches with overlapping degrees between adjacent views. In the test phase, we first estimate or specify the head pose, and then find its suitable view-specific model, and then use the learned cascaded regressor to estimate shapes.

We report the mean error and speed of the baselines and our result for comparison in Tab. 2. Qualitative examples are shown in Fig. 4(a). We can conclude based on the results that the heuristic head pose partition is suboptimal, since it neglects other shape and appearance variations apart from views. It also lacks of robustness from the single cascaded regressor scheme. The multi-view approach also exhibits low speed. The fastest head pose estimation reported in [23] (25 FPS, 10ms/frame with 4 cores) is still much slower than our whole framework. This is because [23] still requires image warping, integral map and gradient calculation, while our approach only requires pixel value look-up from the original image.

4.2. Comparison with existing approaches

4.2.1 In-the-wild datasets (AFLW / AFW)

We compare our approach with various state-of-the-art methods on AFLW and AFW datasets. We exclude the comparison with GSDM [37] on AFLW as it is non-

[7] To our best knowledge, we are the first to evaluate our algorithm on the full set of AFLW. Previous approaches [9, 39, 43, 30] mostly evaluate on a near-frontal subset of AFLW (with $\leq$ 25% samples) due to lack of capability to handle strong shape variation in the full set of AFLW.

[8] We directly use their released codes and settings. This approach achieves state-of-the-art results on the Pointing04 dataset [13].

Table 2. Mean error (% normalised by face size) and speed (FPS) for multi-view approach compared with our framework (Sec. 4.1). We use AFLW-Full setting (Tab. 1). Each baseline is a combination of an head pose estimator and cascaded regressors.

Baseline	[15]+LBF[27]	[38]+LBF[27]	Ground-truth+LBF[27]	[15]+CFSS[44]	[38]+CFSS[44]	Ground-truth+CFSS[44]	**Ours**
Error	3.15	3.19	3.07	3.18	3.23	3.12	**2.72**
FPS	6	24	N.A.	5	13	N.A.	**350**

Table 3. Mean Error (% normalised by face size) on in-the-wild datasets compared with state-of-the-art approaches (Sec. 4.2.1, see Tab. 1 for 4 types of evaluation settings). RCPR, ERT, CFSS (with available training codes) and SDM, LBF (re-implemented) are all reported with results re-trained on the given training set of AFLW. PO-CR is based on the released pre-trained model (on frontal faces) and hence its evaluation hurts on the full-set evaluation (marked with '$*$'). PIFA's result is based on [18].

Evaluation	CDM [40]	RCPR [4]	CFSS [44]	PO-CR [33]	ERT [19]	SDM [36]	LBF [27]	PIFA [18]	**Ours**
AFLW-PIFA	8.59	7.15	6.75	-	7.03	6.96	7.06	6.52	**5.81**
AFLW-Full	5.43	3.73	3.92	5.32^*	4.35	4.05	4.25	-	**2.72**
AFLW-Frontal	3.77	2.87	2.68	2.41	2.75	2.94	2.74	-	**2.17**
AFW	5.70	3.87	3.43	4.37^*	3.25	3.88	3.39	-	**2.45**

applicable to static images. We conduct our experiments for all three types of evaluation (AFLW-PIFA, AFLW-Full, AFLW-Frontal) for AFLW. AFLW-PIFA strictly follows the setting in PIFA [18] and its performance is derived directly from their literature. For AFW, we also follow [18] to use the 3901 training set of AFLW and pick out the 6 landmarks when testing on AFW for evaluation.

Results. Mean error, CED and qualitative examples for both AFLW and AFW are reported in Tab. 3 and Fig. 4. As can be observed from the results, our CCL approach outperforms various approaches by large margin. We note all approaches receive better results in AFLW-Full than AFLW-PIFA because AFLW-PIFA includes vague ear points for evaluation and trained with a smaller training set. Even including ears points in AFLW-PIFA, which is non-favourable for 2D approach, our results still outperform 3D approach (PIFA).

4.2.2 Lab-environment datasets (MultiPIE / FERET)

We conduct comparison on MultiPIE / FERET following the experimental settings of recent studies [34, 17].

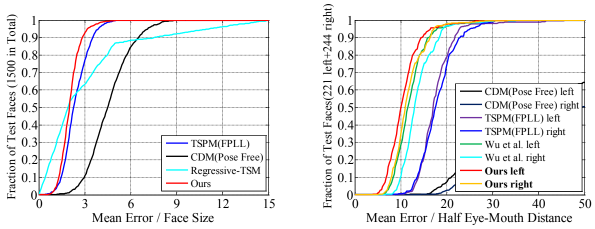
MultiPIE. MultiPIE does not provide train-test partition. We thus follow the evaluation settings in [17] (Tab. 1).

FERET. To compare with [34], we report results on FERET based on the trained model from MultiPIE.

Results. We report our CED curve in Fig. 5. Again, our method outperforms existing approaches though the improvement is less significant compared with that in AFLW/AFW due to the less challenging appearance and shape variations exhibited in the lab environment.

4.3. Ablation Study

Our framework consists of several pivotal components, i.e. robust composition estimation, domain partition, robust feature mapping and on-the-fly domain exclusion. In this section, we validate their effectiveness within our framework. Throughout this section, we use the AFLW-Full for evaluation.



(a) CED for MultiPIE. (b) CED for FERET.

Figure 5. CED for constrained datasets (Sec. 4.2.2).

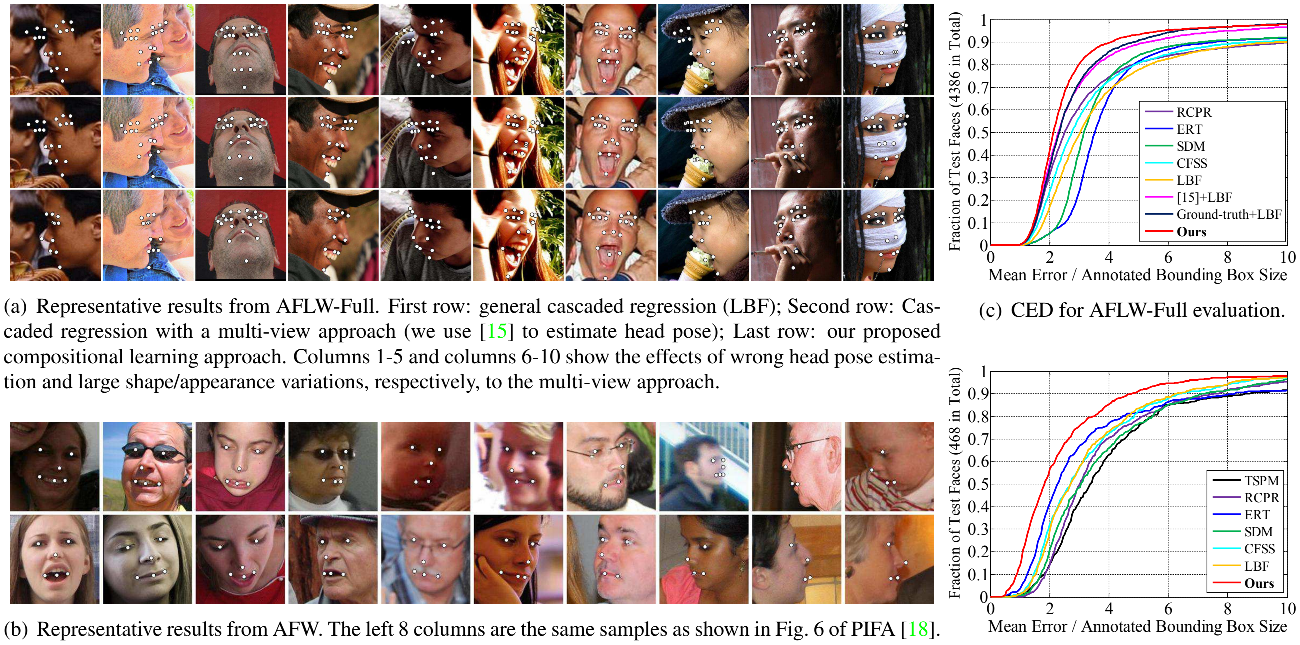
Table 4. Mean error (% normalised by face size) to validate the robust composition estimation. Baseline indicates that we use the indicator vector or classification score from SVM or classification forest.

Baselines	(1)	(2)	(3)	(4)	**Ours**
Error	3.97	4.04	3.86	4.02	**2.72**

Table 5. Mean error (% normalised by face size) for various choices of K in our framework.

K	1	4	16	64
Error	3.73	3.17	2.72	2.69

Validation of robust composition estimation. Estimation of the composition is the core step of our algorithm. The proposed loss function is based on the shape dissimilarity rather than a simple classification loss. We compare against four baselines that predict composition as: (1) SVM's indicator vector; (2) SVM's classification score; (3) classification forest's indicator vector; and (4) classification forest's classification score. We report the mean error of each baseline approach and our result in Tab. 4. The results verify that optimising towards shape rather the indirect domain classification is essential in our framework.



(a) Representative results from AFLW-Full. First row: general cascaded regression (LBF); Second row: Cascaded regression with a multi-view approach (we use [15] to estimate head pose); Last row: our proposed compositional learning approach. Columns 1-5 and columns 6-10 show the effects of wrong head pose estimation and large shape/appearance variations, respectively, to the multi-view approach.

(c) CED for AFLW-Full evaluation.

(b) Representative results from AFW. The left 8 columns are the same samples as shown in Fig. 6 of PIFA [18]. The last two columns are more selected challenging samples.

(d) CED for AFW evaluation.

Figure 4. (a) Representative examples for comparison on cascaded regression, multi-view approach and our compositional learning method on AFLW-Full; (b) Comparison between our approach with PIFA [18] by choosing the same examples as in [18]; (c-d) CED curve for AFLW-Full and AFW (Sec. 4.1 and Sec. 4.2.1).

Validation of DHD partition. Partitioning the optimisation space into domains of homogeneous descent plays vital role in our framework. We verify the effectiveness of the partition by evaluating accuracy with various choices of K (number of domains). As shown in Tab. 5, the accuracy improves dramatically with K rising from 1 to 4. A K larger than 16 leads to marginal improvement. The results suggest the importance of domain partition, and also justify the choice of $K = 16$ in our experiments.

Validation of the robust feature mapping. Robust feature mapping not only provides robust feature encoding to cope with the self-occlusion issue inherent in the unconstrained scenario, but also provides essential information to facilitate composition prediction. Without the additional visibility encoding, the mean error drops from 2.72 to 3.78.

Validation of on-the-fly domain exclusion. We introduce this strategy to speed up the inference of our framework. Improvement is negligible if we do not exclude any domain among iterations. However, without domain exclusion, the speed would reduce to 90 FPS, which is roughly 4 times slower than the setting with exclusion enabled (350 FPS).

5. Conclusion

Unconstrained face alignment is an emerging topic. In this paper we have presented a novel and practical method to this problem. By estimating a shape as a composition of various domains of homogeneous descent, the method is capable of handling arbitrary head poses as well as large shape and appearance variations. Owing to on-the-fly domain exclusion with fast and robust feature mapping, our method achieves accurate compositional shape estimation and high inference speed.

Acknowledgement

This work is partially supported by SenseTime Group Limited and the Hong Kong Innovation and Technology Support Programme.

References

[1] J. Alabort-i Medina and S. Zafeiriou. Unifying holistic and parts-based deformable model fitting. In *CVPR*, 2015. 1

[2] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, pages 545–552, 2011. 6

[3] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 5

[4] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *ICCV*, pages 1513–1520, 2013. 5, 7

[5] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *IJCV*, 107(2):177–190, 2014. 1, 2

[6] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In *ECCV*, pages 109–122, 2014. 5

[7] G. G. Chrysos, E. Antonakos, P. Snape, A. Asthana, and S. Zafeiriou. A comprehensive performance evaluation of deformable face tracking" in-the-wild". *arXiv preprint arXiv:1603.06015*, 2016. 1

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *TPAMI*, 23(6):681–685, 2001. 1

[9] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. In *ECCV*. 2012. 6

[10] A. Datta, R. Feris, and D. Vaquero. Hierarchical ranking of facial attributes. In *AFGR*, pages 36–42. IEEE, 2011. 1

[11] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *CVPR*, pages 1078–1085, 2010. 1

[12] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *CVPR*, 2009. 5

[13] N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial structures. In *FG Net Workshop on Visual Observation of Deictic Gestures*, 2004. 6

[14] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 2010. 6

[15] K. Hara and R. Chellappa. Growing regression forests by classification: Applications to object pose estimation. In *ECCV*. 2014. 1, 6, 7, 8

[16] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *CVPR*, 2015. 1

[17] G.-S. Hsu, K.-H. Chang, and S.-C. Huang. Regressive tree structured model for facial landmark localization. In *ICCV*, 2015. 1, 2, 6, 7

[18] A. Jourabloo and X. Liu. Pose-invariant 3d face alignment. In *ICCV*, 2015. 1, 2, 6, 7, 8

[19] V. Kazemi and S. Josephine. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014. 7

[20] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCVW*, 2011. 1, 2, 5

[21] H. Lai, S. Xiao, Z. Cui, Y. Pan, C. Xu, and S. Yan. Deep cascaded regression for face alignment. *arXiv preprint arXiv:1510.09083*, 2015. 1

[22] D. Lee, H. Park, and C. D. Yoo. Face alignment using cascade gaussian process regression trees. In *CVPR*, 2015. 5

[23] D. Lee, M.-H. Yang, and S. Oh. Fast and accurate head pose estimation via random projection forests. In *ICCV*, 2015. 2, 6

[24] D. Miller, I. Kemelmacher-Shlizerman, and S. M. Seitz. Megaface: A million faces for recognition at scale. *arXiv preprint arXiv:1505.02108*, 2015. 1

[25] X. Peng, S. Zhang, Y. Yu, and D. N. Metaxas. Piefa: Personalized incremental and ensemble face alignment. In *ICCV*, 2015. 1

[26] P. J. Phillips, H. Moon, S. Rizvi, P. J. Rauss, et al. The feret evaluation methodology for face-recognition algorithms. *TPAMI*, 2000. 6

[27] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *CVPR*, 2014. 1, 3, 4, 5, 6, 7

[28] 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. 1, 5

[29] 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 *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 50–58, 2015. 1

[30] X. Shen, Z. Lin, J. Brandt, and Y. Wu. Detecting and aligning faces by image retrieval. In *CVPR*, 2013. 6

[31] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, pages 3476–3483, 2013. 1

[32] S. Tulyakov and N. Sebe. Regressing a 3d face shape from a single image. In *ICCV*, 2015. 1, 2

[33] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *CVPR*, 2015. 1, 5, 7

[34] Y. Wu and Q. Ji. Robust facial landmark detection under significant head poses and occlusion. In *ICCV*, 2015. 1, 2, 5, 6, 7

[35] S. Xiao, S. Yan, and A. Kassim. Facial landmark detection via progressive initialization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 33–40, 2015. 1

[36] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013. 1, 2, 3, 7

[37] X. Xiong and F. De la Torre. Global supervised descent method. In *CVPR*, 2015. 1, 2, 3, 4, 6

[38] H. Yang, W. Mou, Y. Zhang, I. Patras, H. Gunes, and P. Robinson. Face alignment assisted by head pose estimation. In *BMVC*, 2015. 6, 7

[39] H. Yang and I. Patras. Sieving regression forest votes for facial feature detection in the wild. In *ICCV*, 2013. 6

[40] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *ICCV*, 2013. 6, 7

[41] X. Yu, Z. Lin, J. Brandt, and D. N. Metaxas. Consensus of regression for occlusion-robust facial feature localization. In *ECCV*. 2014. 5

[42] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *Computer Vision–ECCV 2014*, pages 94–108. Springer, 2014. 1

[43] X. Zhao, T.-K. Kim, and W. Luo. Unified face analysis by iterative multi-output random forests. In *CVPR*, pages 1765–1772, 2014. 6

[44] S. Zhu, C. Li, C. C. Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, 2015. 5, 6, 7

[45] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886, 2012. 2, 5, 6