

Structured Landmark Detection via Topology-Adapting Deep Graph Learning

Weijian Li^{1,2}, Yuhang Lu^{1,3}, Kang Zheng¹, Haofu Liao², Chihung Lin⁴, Jiebo Luo², Chi-Tung Cheng⁴, Jing Xiao⁵, Le Lu¹, Chang-Fu Kuo⁴, and Shun Miao¹

¹ PAII, Inc., Bethesda, MD, USA

² Department of Computer Science, University of Rochester, Rochester, NY, USA

³ Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA

⁴ Chang Gung Memorial Hospital, Linkou, Taiwan, ROC

⁵ Ping An Technology, Shenzhen, China

Abstract. Image landmark detection aims to automatically identify the locations of predefined fiducial points. Despite recent success in this field, higher-ordered structural modeling to capture implicit or explicit relationships among anatomical landmarks has not been adequately exploited. In this work, we present a new topology-adapting deep graph learning approach for accurate anatomical facial and medical (e.g., hand, pelvis) landmark detection. The proposed method constructs graph signals leveraging both local image features and global shape features. The adaptive graph topology naturally explores and lands on task-specific structures which are learned end-to-end with two Graph Convolutional Networks (GCNs). Extensive experiments are conducted on three public facial image datasets (*WFLW*, *300W*, and *COFW-68*) as well as three real-world X-ray medical datasets (*Cephalometric (public)*, *Hand* and *Pelvis*). Quantitative results comparing with the previous state-of-the-art approaches across all studied datasets indicating the superior performance in both robustness and accuracy. Qualitative visualizations of the learned graph topologies demonstrate a physically plausible connectivity laying behind the landmarks.

Keywords: Landmark Detection, GCN, Adaptive Topology

1 Introduction

Image landmark detection has been a fundamental step for many high-level computer vision tasks to extract and distill important visual contents, such as image registration [23], pose estimation [4], identity recognition [77] and image super-resolution [5]. Robust and accurate landmark localization becomes a vital component determining the success of the downstream tasks.

Recently, heatmap regression based methods [62,74,49,55] have achieved encouraging performance on landmark detection. They model landmark locations as heatmaps and train deep neural networks to regress the heatmaps. Despite popularity and success, they usually suffer from a major drawback of lacking a

global representation for the structure/shape, which provides high-level and reliable cues in individual anatomical landmark localization. As a result, heatmap-based methods could make substantial errors when being exposed to large appearance variations such as occlusions.

In contrast, coordinate regression based methods [36, 71, 69, 54] have an innate potential to incorporate structural knowledge since the landmark coordinates are directly expressed. Most existing methods initialize landmark coordinates using mean or canonical shapes, which indirectly inject weak structural knowledge [54]. While the exploitation of the structural knowledge in existing methods has still been insufficient as well as further exploitation of the structural knowledge considering the underlying relationships between the landmarks. Effective means for information exchange among landmarks to facilitate landmark detection are also important but have yet to be explored. Due to these limitations, the performance of the latest coordinate-based methods [63] falls behind the heatmap-based ones [58].

In this work, we introduce a new topology-adapting deep graph learning approach for landmark detection, termed *Deep Adaptive Graph (DAG)*. We model the landmarks as a graph and employ global-to-local cascaded Graph Convolutional Networks (GCNs) to move the landmarks towards the targets in multiple steps. Graph signals of the landmarks are built by combining local image features and graph shape features. Two GCNs operate in a cascaded manner, with the first GCN estimating a global transformation of the landmarks and the second GCN estimating local offsets to further adjust the landmark coordinates. The graph topology, represented by the connectivity weights between landmarks, are learned during the training phase.

By modeling landmarks as a graph and processing it with GCNs, our method is able to effectively exploit the structural knowledge and allow rich information exchange among landmarks for accurate coordinate estimation. The graph topology learned for landmark detection task is capable of revealing reasonable landmark relationships for the given task. It also reduces the need for manually defining landmark relations (or grouping), making our method to be easily adopted for different tasks. By incorporating shape features into graph signal in addition to the local image feature, our model can learn and exploit the landmark shape prior to achieve high robustness against large appearance variations (e.g., occlusions). In summary, our main contributions are four-fold:

1. By representing the landmarks as a graph and detecting them using GCNs, our method effectively exploits the structural knowledge for landmark coordinate regression, closes the performance gap between coordinate- and heatmap-based landmark detection methods.
2. Our method automatically reveals physically meaningful relationships among landmarks, leading to a task-agnostic solution for exploiting structural knowledge via step-wise graph transformations.
3. Our model combines both visual contextual information and spatial positional information into the graph signal, allowing structural shape prior to be learned and exploited.

4. Comprehensive quantitative evaluations and qualitative visualizations on six datasets across both facial and medical image domains demonstrate the consistent state-of-the-art performance and general applicability of our method.

2 Related Work

A large number of studies have been reported in this domain including the classic Active Shape Models [37,13,12], Active Appearance Models [11,46,33], Constraint Local Models [14,3,45,31], and more recently the deep learning based models which can be further categorized into heatmap or regression based models.

Heatmap Based Landmark Detection: These methods [60,38,51,49,39,10] generate localized predictions of likelihood heatmaps for each landmark and achieve encouraging performances. A preliminary work by Wei *et al.* [60] introduce a Convolutional Pose Machine (CPM) which models the long-range dependency with a multistage network. Newell *et al.* [38] propose a Stacked Hourglass model leveraging the repeated bottom-up and top-down structure and intermediate supervision. Tang *et al.* [51] investigate a stacked U-Net structure with dense connections. Lately, Sun *et al.* [49] present a deep model named High-Resolution Network (HRNet18) which extracts feature maps in a joint deep and high resolution manner via conducting multi-scale fusions across multiple branches under different resolutions. Based on these models, other methods also integrate additional supervision cues such as the object structure constraints [64,78], the variety of image, and object styles [18,42] to solve specific tasks.

Coordinate Based Landmark Detection: Another common approach directly locates landmark coordinates from input images [53,50,54,36,75,34,48]. Most of these methods consist of multiple steps to progressively update predictions based on visual signals, widely known as Cascaded-Regression. Toshev *et al.* [53] and Sun *et al.* [50] adopt cascaded Convolutional Neural Networks (CNNs) to predict landmark coordinates. Trigeorgis *et al.* [54] model the cascaded regression process using a Recurrent Neural Network (RNN) based deep structure. Lv *et al.* [36] propose a two-stage regression model with global and local reinitializations. From different perspectives, Zhu *et al.* [75] investigate the methods of optimal initialization by searching the object shape space; Valle *et al.* [55] present a combined model with a tree structured regressor to infer landmark locations based on heatmap prediction results; Wu *et al.* [63] leverage uniqueness and discriminative characteristics across datasets to assist landmark detection.

Landmark Detection with Graphs: The structure of landmarks can be naturally modeled as a graph considering the landmark locations and landmark to landmark relationships [73,67,45,68,75]. Zhou *et al.* [73] propose a Graph-Matching method which obtains landmark locations by selecting the set of landmark candidates that would best fit the shape constraints learned from the exemplars. Yu *et al.* [68] describe a two-stage deformable shape model to first extract a coarse optimum by maximizing a local alignment likelihood in the region of interest then refine the results by maximizing an energy function under

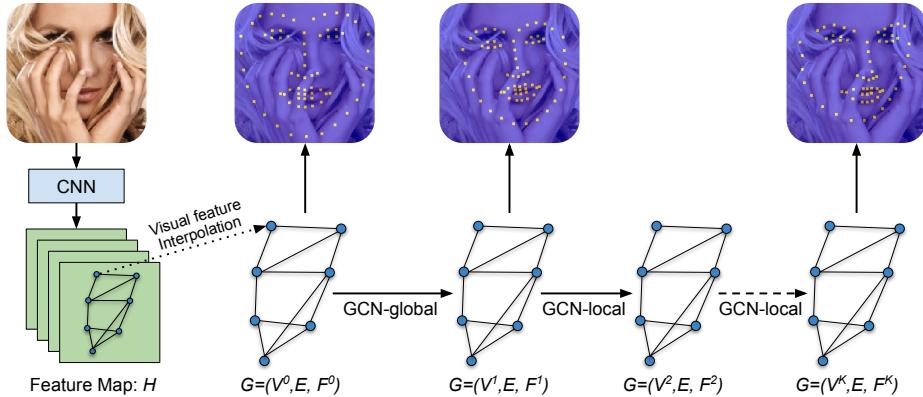


Fig. 1: Overview of the proposed Deep Adaptive Graph (DAG). Initial graph is initialized with the mean value computed from training data. We first deform the landmark graph through a perspective transformation predicted by GCN-global and then precisely shift the graph by GCN-local through iterations. The visual features and shape features are re-interpolated from feature map and re-calculated after each GCN module, respectively.

shape constraints. Later, Yu *et al.* [67] present a hierarchical model to extract semantic features by constructing intermediate graphs from bottom-up node clustering and top-down graph deconvolution operations, leveraging the graph layout information. Zou *et al.* [78] introduce a landmark structure construction method with covering set algorithm. While their method is based on heatmap detection results, we would like to directly regress landmark locations from raw input image to avoid potential errors incurred from heatmap detections.

Recently, Ling *et al.* [32] propose a fast object annotation framework, where contour vertices are regressed using GCN to perform segmentation, indicating the benefit of position prediction with iterative message exchanges. In their task, each point is considered with the same semantics towards coarse anonymous matching which is not appropriate for precise targeted localization tasks like landmark detection. Adaptively learning graph connectivities instead of employing a fixed graph structure based on prior knowledge should be explored to improve the model’s generalizability to different tasks.

3 Method

Our method adopts the cascaded-regression framework, where given the input image and initial landmarks (from the mean shape), the predicted landmark coordinates are updated in multiple steps. Yet differently, we feature the cascaded-regression framework with a graph representation of the landmarks, denoted by $G = (V, E, F)$, where $V = \{\mathbf{v}_i\}$ denotes the landmarks, $E = \{e_{ij}\}$ denotes the learned connectivity between landmarks and $F = \{\mathbf{f}_i\}$ denotes graph signals

capturing appearance and shape information. The graph is processed by cascaded GCNs to progressively update landmark coordinates. An overview of our method is shown in Figure 1. Details of the cascaded GCNs, graph signal and learned connectivity are presented in Section 3.1, Section 3.2 and Section 3.3, respectively. The training scheme of our method can be found in Section 3.4.

3.1 Cascaded GCNs

Given a graph representation of landmarks $G = (V, E, F)$, two-stage cascaded GCN modules are employed to progressively update the landmark coordinates. The first stage, *GCN-global*, estimates a global transformation to coarsely move the landmarks to the targets. The second stage, *GCN-local*, estimates local landmark coordinate offsets to iteratively move the landmarks toward the targets. Both modules employ the same GCN architecture (weights not shared) and the same learnable graph connectivity.

Graph Convolution: Given a graph connectivity E and a graph feature F , the k -th graph convolution operation updates the i -th node feature \mathbf{f}_k^i by aggregating all node features weighted by the connectivity:

$$\mathbf{f}_{k+1}^i = \mathbf{W}_1 \mathbf{f}_k^i + \sum_j e_{ij} \mathbf{W}_2 \mathbf{f}_k^j \quad (1)$$

where \mathbf{W}_1 and \mathbf{W}_2 are learnable weight matrices. The graph convolutions can be seen as the mechanism of information collection among the neighborhoods. The connectivity E serves as pathways for information flow from one landmark to another.

Global Transformation GCN: Previous work [27,36] learn an affine transformation with a deep neural network by predicting a two by three affine transformation matrix which deforms the image to the satisfied posture. Inspired by this work, we employ a GCN on the initial landmarks to coarsely move them to the targets. Considering our graph is more flexible than does not have to maintain the parallelism and respective ratios among the edges, we model the global transformation using a perspective transformation [17]. A perspective transformation can be parameterized by 9 scalars $M = [a, b, c, d, e, f, g, h, i]^T \in \mathbb{R}^{9 \times 1}$ with the operation written as:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} \cong \begin{bmatrix} rx' \\ ry' \\ r \end{bmatrix} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (2)$$

Given a target image, we initialize landmark locations V^0 using the mean shape of landmarks in the training set, and placed it at the center of the image. The graph is processed by the GCN-global to estimate a perspective transformation to bring the initial structure closer to the target.

Specifically, a graph isomorphism network (GIN) [66] is employed to process the graph features $\{\mathbf{f}_k^i\}$ produced by the GCN to output a 9-dimensional vector

representing the perspective transformation:

$$\mathbf{f}^G = \text{MLP}(\text{CONCAT}(\text{READOUT}(\{\mathbf{f}_k^i | i \in G\}) | k = 0, 1, \dots, K)), \quad (3)$$

where the READOUT operator sums the features from all the nodes in the graph G . The transformation matrix M is obtained by transforming and reshaping \mathbf{f}^G into a 3 by 3 matrix. We then apply this transformation matrix on the initial landmark node coordinates to obtain the aligned landmark coordinates:

$$V^1 = \{\mathbf{v}_i^1\} = \{\mathbf{M}\mathbf{v}_i^0\} \quad (4)$$

Local Refinement GCN: Given the transformed landmarks, we employ GCN-local to further shift the graph in a cascaded manner. GCN-local employs the same architecture as GCN-global, with a difference that the last layer produces a 2-dimensional vector for each landmark, representing the coordinate offset of the landmark. The updated landmark coordinates can be written as:

$$\mathbf{v}_i^{t+1} = \mathbf{v}_i^t + \Delta\mathbf{v}_i^t, \quad (5)$$

where $\Delta\mathbf{v}_i^t = (\Delta x_i^t, \Delta y_i^t)$ is the output of the GCN-local at the t -th step. In all our experiments, we perform $T = 3$ iterations of the GCN-local. Note that the graph signal is re-calculated after each GCN-local iteration.

3.2 Graph signal with appearance and shape information

We formulate a graph signal F as a set of node features \mathbf{f}_i , each associated with a landmark \mathbf{v}_i . The graph signal contains a *visual feature* to encode local image appearance and a *shape feature* to encode the global landmark shape.

Visual Feature: Specifically, given a feature map H with D channels produced by a backbone CNN, visual features, denoted by $\mathbf{p}_i \in R^D$, are extracted by interpolating H at the landmark coordinates \mathbf{v}_i . The interpolation is performed via a differentiable bi-linear interpolation [27]. In this way, visual feature of each landmark is collected from the feature map, encoding the appearance of its neighborhood.

Shape Feature: While the visual feature encodes the appearance in a neighborhood of the landmark, it does not explicitly encode the global shape of the landmarks. To incorporate this structural information into the graph signal, for each landmark, we compute its displacement vectors to all other landmarks, denoted as $\mathbf{q}_i = \{\mathbf{v}_j - \mathbf{v}_i\}_{j \neq i} \in R^{2 \times (N-1)}$, where N is the number of landmarks. Such shape feature allows structural information of the landmarks to be exploited to facilitate landmark detection. For example, when the mouth of a face is occluded, the coordinates of the mouth landmarks can be inferred from the eyes and nose. Wrong landmark detection results that violate the shape prior can also be avoided when the shape is explicitly captured in the graph signal.

The graph signal F is then constructed for each landmark by concatenating the visual feature \mathbf{p}_i and the shape feature \mathbf{q}_i (flattened), resulting in a feature vector $\mathbf{f}_i \in R^{D+2(N-1)}$.

3.3 Landmark graph with learnable connectivity

The graph connectivity determines the relationship between each pair of landmarks in the graph and serves as the information exchange channel in GCN. In most existing applications of GCN [41,32,72,56,61], the graph connectivity is given based on the prior knowledge of the task. In our landmark detection application, it is non-trivial to manually define the optimal underlying graph connectivity for the learning task. Therefore, relying on hand-crafted graph connectivity would introduce a subjective element into the model, which could lead to sub-optimal performance. To address this limitation, we learn task-specific graph connectivities during the training phase in an end-to-end manner. The connectivity weight e_{ij} behaves as information propagation gate in graph convolutions (Eqn. 1). We treat the connectivity $\{e_{ij}\}$, represented as an adjacency matrix, as a learnable parameter that is trained with the network during the training phase. In this way, the task-specific optimal graph connectivity is obtained by optimizing the performance of the target landmark detection task, allowing our method to be applied to different landmark detection tasks without manual intervention.

Graph connectivity learning has been studied before by the research community. One notable example is Graph Attention Networks [56], which employs a self-attention mechanism to adaptively generate connectivity weights during the model inference. We conjugate that in structured landmark detection problems, the underlying relationship between the landmarks remains the same for a given task, instead of varying across individual images. Therefore, we share the same connectivity across images on the same task, and directly optimize the connectivity weights during the training phase.

3.4 Training

GCN-global: Since the perspective transformation estimated by GCN-global has limited degree of freedom, directly penalizing the distance between the predicted and the ground truth landmarks will lead to unstable optimization behavior. As the goal of GCN-global is to coarsely locate the landmarks, we propose to use a margin loss on the L_1 distance, written as:

$$\mathcal{L}_{global} = \left[\left(\frac{1}{N} \sum_{i \in N} \sum_{x,y} |\mathbf{v}_i^1 - \mathbf{v}_i| \right) - m \right]_+ \quad (6)$$

where $[u]_+ := \max(0, u)$. $\mathbf{v}_i^1 = (x_i^1, y_i^1)$ and $\mathbf{v}_i = (x_i, y_i)$ denote the predicted and ground truth landmark coordinates for the i -th landmark. m is a hyper-parameter representing a margin which controls how well we want the alignment to be. Following this procedure, we aim to obtain a high robustness of the coarse landmark detection, while forgive small errors.

GCN-local: To learn a precise localization, we directly employ L1 loss on all predicted landmark coordinates after the GCN-local, written as:

$$\mathcal{L}_{local} = \frac{1}{N} \sum_{i \in N} \sum_{x,y} |\mathbf{v}_i^T - \mathbf{v}_i| \quad (7)$$

where \mathbf{v}_i^T is the T -th step (the last step) coordinate predictions, and \mathbf{v}_i is the ground truth coordinate for the i -th landmark.

The overall loss to train DAG is a combination of the above two losses:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{global} + \lambda_2 \mathcal{L}_{local} \quad (8)$$

where λ_k is the weight parameter for each loss.

4 Experiments

4.1 Datasets

We conduct evaluations on three public facial image and three medical image datasets:

WFLW [62] dataset contains 7,500 facial images for training and 2,500 facial images for testing. The testing set is further divided into 6 subsets focusing on particular challenges in the images namely large pose set, expression set, illumination set, makeup set, occlusion set, and blur set. 98 manually labeled landmarks are provided for each image.

300W [44] dataset consists of 5 facial datasets namely LFPW, AFW, HELEN, XM2VTS and IBUG. They are split into a training set with 3,148 images, and a testing set with 689 images where 554 images are from LFPW and HELEN, 135 from IBUG. Each image is labeled with 68 landmarks.

COFW [6] dataset contains 1,345 facial images for training and 507 for testing, under different occlusion conditions. Each image is originally labeled with 29 landmarks and re-annotated with 68 landmarks [22]. We follow previous studies [62,42] to conduct inferences on the re-annotated COFW-68 dataset to test our model’s cross-dataset performance which is trained on 300W dataset.

Cephalometric X-ray [57] is a public dataset originally for a challenge in IEEE ISBI-2015. It contains 400 X-ray Cephalometric images with resolution of $1,935 \times 2,400$, 150 images are used as training set, the rest 150 images and 100 images are used as validation and test sets. Each cephalometric image contains 19 landmarks. In this paper, we only focus on the landmark detection task.

Hand X-ray [35] is a real-world medical dataset collected by a hospital. The X-ray images are taken with different hand poses with resolutions in $1,500s \times 2,000s$. In total, 471 images are randomly split into a training set (80%, $N=378$) and a testing set (20%, $N=93$). 30 landmarks are manually labeled for each image.

Pelvic X-ray [59,9] another real-world medical dataset collected by the same hospital. Images are taken over patient’s pelvic bone with resolutions in $2,500s \times 2,000s$. The challenges in this dataset is the high structural and appearance variation, caused by bone fractures and metal prosthesis. In total, 1,000 images are randomly split into a training set (80%, $N=800$) and a testing set (20%, $N=200$). 16 landmarks are manually labeled for each image.

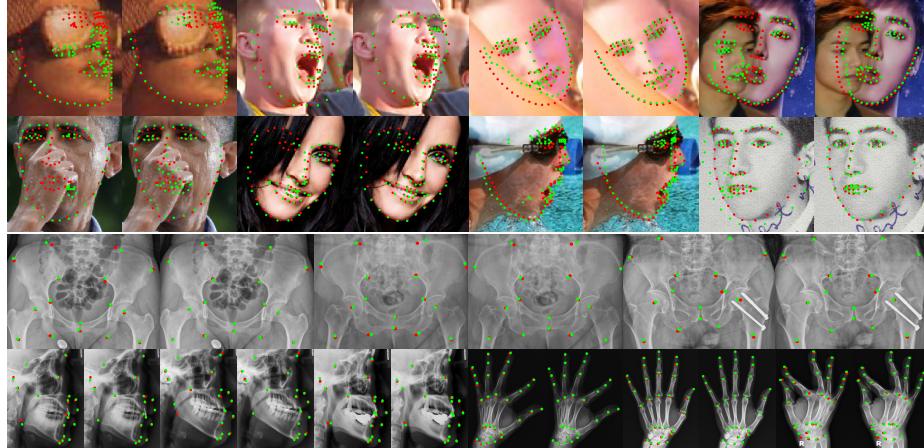


Fig. 2: Visualization of landmark detection results. Pairs of results are displayed side by side. For each pair, **Left image**: detection result from a SOTA method [49]. **Right image**: result produced by our method. **Green dot**: predicted landmark location. **Red dot**: groundtruth landmark location.

4.2 Experiment Settings

Evaluation Metrics: We evaluate the proposed method following two sets of metrics. For the facial image datasets, we employ the widely adopted Normalized Mean Error (NME), Area Under the Curve (AUC), Failure Rate for a maximum error of 0.1 (FR@0.1) and Cumulative Errors Distribution (CED) curve (supplementary material). To compare with previous methods, we conduct both "interocular" (outer-eye-corner-distance) and "inter-pupil" (eye-center-distance) normalizations on the detected landmark coordinates.

For the Cephalometric X-ray images, we follow the original evaluation protocol to compare two sets of metrics: Mean Radial Error (MRE) which computes the average of Euclidean Distances of predicted coordinates and ground truth coordinates of all the landmarks; the corresponding Successful Detection Rate (SDR) under 2mm, 2.5mm, 3mm and 4mm. For the Hand and Pelvic X-rays, we compute MRE, Hausdorff Distance (HD) and Standard Deviations (STD). Recall that Hausdorff Distance measures the maximum value of the minimum distances between two sets of points. In our case, we aim to evaluate the error upper-bound for the detected landmarks.

Implementation Details: Following previous studies, we crop and resize facial images into 256×256 based on the provided bounding boxes. We follow [10] to resize the Cephalometric X-rays to 640×800 . For the Hand and Pelvic X-rays, we resize each image into 512×512 preserving the original height and width ratio by padding zero values to the empty regions. The proposed model is implemented in PyTorch and is experimented on a single NVIDIA Titan V GPU. We choose $\lambda_1 = \lambda_2 = 1$ for different parts in the overall loss function. HRNet18 [49] pretrained on

Table 1: Evaluation on the WFLW dataset (98 Landmarks). *: focus on loss function. #: focus on data augmentation.

Metric	Method	Test	Pose	Expression	Illumination	Make-up	Occlusion	Blur
Mean Error %	CFSS [75]	9.07	21.36	10.09	8.30	8.74	11.76	9.96
	DVLN [63]	6.08	11.54	6.78	5.73	5.98	7.33	6.88
	LAB [62]	5.27	10.24	5.51	5.23	5.15	6.79	6.32
	SAN [18] #	5.22	10.39	5.71	5.19	5.49	6.83	5.80
	WING [21] *	5.11	8.75	5.36	4.93	5.41	6.37	5.81
	HRNet18 [49]	4.60	7.94	4.85	4.55	4.29	5.44	5.42
	STYLE [42] #	4.39	8.42	4.68	4.24	4.37	5.60	4.86
Failure Rate @0.1	AWING [58] *	4.36	7.38	4.58	4.32	4.27	5.19	4.96
	Ours	4.21	7.36	4.49	4.12	4.05	4.98	4.82
	CFSS [75]	20.56	66.26	23.25	17.34	21.84	32.88	23.67
Failure Rate @0.1	DVLN [63]	19.84	46.93	11.15	7.31	11.65	16.30	13.71
	LAB [62]	7.56	28.83	6.37	6.73	7.77	13.72	10.74
	SAN [18] #	6.32	27.91	7.01	4.87	6.31	11.28	6.60
	WING [21] *	6.00	22.70	4.78	4.30	7.77	12.50	7.76
	HRNet18 [49]	4.64	23.01	3.50	4.72	2.43	8.29	6.34
	STYLE [42] #	4.08	18.10	4.46	2.72	4.37	7.74	4.40
	AWING [58] *	2.84	13.50	2.23	2.58	2.91	5.98	3.75
AUC @0.1	Ours	3.04	15.95	2.86	2.72	1.45	5.29	4.01
	CFSS [75]	0.3659	0.0632	0.3157	0.3854	0.3691	0.2688	0.3037
	DVLN [63]	0.4551	0.1474	0.3889	0.4743	0.4494	0.3794	0.3973
	HRNet18 [49]	0.5237	0.2506	0.5102	0.5326	0.5445	0.4585	0.4515
	LAB [62]	0.5323	0.2345	0.4951	0.5433	0.5394	0.4490	0.4630
	SAN [18] #	0.5355	0.2355	0.4620	0.5552	0.5222	0.4560	0.4932
	WING [21] *	0.5504	0.3100	0.4959	0.5408	0.5582	0.4885	0.4932
AUC @0.1	AWING [58] *	0.5719	0.3120	0.5149	0.5777	0.5715	0.5022	0.5120
	STYLE [42] #	0.5913	0.3109	0.5490	0.6089	0.5812	0.5164	0.5513
	Ours	0.5893	0.3150	0.5663	0.5953	0.6038	0.5235	0.5329

ImageNet is used as our backbone network to extract visual feature maps for its parallel multi-resolution fusion mechanism and deep network design which fits our need for both high resolution and semantic feature representation. The last output after fusion is extracted as feature map of dimension $H \in R^{256 \times 64 \times 64}$. We employ 4 residual GCM blocks [32,30] in GCN-global and GCN-local and perform 3 iterations of GCN-local. Adjacency matrix values are initialized to $1/N$ so that the total weight for each node is 1 to avoid message explosion.

4.3 Comparison with the SOTA methods

WFLW: WFLW is a comprehensive public facial landmark detection dataset focusing on multi-discipline and difficult detection scenarios. Summary of results is shown in Table 1. Following previous works, three evaluation metrics are computed: Mean Error, FR@0.1 and AUC@0.1. Our model achieves 4.21% mean error which outperforms all the strong state-of-the-art methods including AWING [58] which adopts a new adaptive loss function, SAN [18] and STYLE [42] which leverage additional generated images for training. The most significant improvements lie in *Make-up* and *Occlusion* subsets, where only partial landmarks are visible. Our model is able to accurately infer those hard cases based on the visible landmarks due to the benefit of preserving and leveraging graph structural knowledge. This can be further illustrated by examining the visualization results for the occlusion scenarios in Figure 2.

Table 2: Evaluation on 300W Common set, Challenge set and Fullset.

Inter-Pupil Normalization					
Method	Year	Comm.	Challenge	Full.	
CFAN [70]	2014	5.50	16.78	7.69	
ESR [7]	2014	5.28	17.00	7.58	
SDM [65]	2013	5.57	15.40	7.52	
3DDFA [76]	2016	6.15	10.59	7.01	
LBF [43]	2014	4.95	11.98	6.32	
CFSS [75]	2015	4.73	9.98	5.76	
SeqMT [25]	2018	4.84	9.93	5.74	
TCDCN [71]	2015	4.80	8.60	5.54	
RCN [26]	2016	4.67	8.44	5.41	
TSR [36]	2017	4.36	7.56	4.99	
DVLN [63]	2017	3.94	7.62	4.66	
HG-HSLE [78]	2019	3.94	7.24	4.59	
DCFE [55]	2018	3.83	7.54	4.55	
STYLE [42] #	2019	3.98	7.21	4.54	
AWING [58] *	2019	3.77	6.52	4.31	
LAB [62]	2018	3.42	6.98	4.12	
WING [21] *	2018	3.27	7.18	4.04	
Ours	2020	3.64	6.88	4.27	
Inter-Ocular Normalization					
Method	Year	Comm.	Challenge	Full.	
PCD-CNN [29]	2018	3.67	7.62	4.44	
ODN [74]	2019	3.56	6.67	4.17	
CPM+SBR [19]	2018	3.28	7.58	4.10	
SAN [18] #	2018	3.34	6.60	3.98	
STYLE [42] #	2019	3.21	6.49	3.86	
LAB [62]	2018	2.98	5.19	3.49	
HRNet18 [49]	2019	2.91	5.11	3.34	
HG-HSLE [78]	2019	2.85	5.03	3.28	
LUVLi [28]	2020	2.76	5.16	3.23	
AWING [58] *	2019	2.72	4.52	3.07	
Ours	2020	2.62	4.77	3.04	

Table 3: Evaluation on 300W and COFW-68 testsets with the model trained on 300W training set.

300W			
Method	Year	AUC@0.1	FR@0.1
Deng <i>et al.</i> [15]	2016	0.4752	5.50
Fan <i>et al.</i> [20]	2016	0.4802	14.83
DensReg+DSM [1]	2017	0.5219	3.67
JMFA [16]	2019	0.5485	1.00
LAB [62]	2018	0.5885	0.83
HRNet18 [49]	2019	0.6041	0.66
AWING [58] *	2019	0.6440	0.33
Ours	2020	0.6361	0.33
COFW-68			
Method	Year	Mean Error %	FR@0.1
CFSS [75]	2015	6.28	9.07
HRNet18 [49]	2019	5.06	3.35
LAB [62]	2018	4.62	2.17
STYLE [42] #	2019	4.43	2.82
Ours	2020	4.22	0.39

Table 4: Evaluations on the hand X-ray and pelvic X-ray images.

Hand X-ray Dataset				
Method	Year	MRE (pix)	Hausdorff	STD
HRNet18 [49]	2019	12.79	26.36	6.07
Chen <i>et al.</i> [10]	2019	7.14	18.71	14.43
Payer <i>et al.</i> [40]	2019	6.11	16.55	4.01
Ours	2020	5.57	14.83	3.63
Pelvic X-ray Dataset				
Method	Year	MRE (pix)	Hausdorff	STD
HRNet18 [49]	2019	24.77	71.31	19.98
Payer <i>et al.</i> [40]	2019	20.96	68.19	21.93
Chen <i>et al.</i> [10]	2019	20.10	59.92	20.14
Ours	2020	18.39	56.72	17.67

300W: There are two evaluation protocols, namely inter-pupil and inter-ocular normalizations. In this paper, we conduct experiments under both settings on the detection results in order to comprehensively evaluate with the other state-of-the-arts. As can be seen from Table 2, our model achieves competitive results in both evaluation settings comparing to the previous best models, STYLE [42], LAB [62] and AWING [58] which are all heatmap-based. Comparing to the latest coordinate-based model ODN [74] and DVLN [63], our method achieves improvements in large margins (27% and 8% respectively) which sets a remarkable milestone for coordinate-based models, closing the gap between coordinate- and heatmap-based methods.

COFW-68 and 300W testset: To verify the robustness and generalizability of our model, we conduct inference on images from COFW-68 and 300W testset using the model trained on 300W training set and validated on 300W fullset. Results summarized in Table 3 indicating our model’s superior performance over

Table 5: Evaluation on the public Cephalometric dataset.

Model	Year	Validation set				Test set					
		MRE	2mm	2.5mm	3mm	4mm	MRE	2mm	2.5mm	3mm	4mm
Arik <i>et al.</i> [2]	2017	-	75.37	80.91	84.32	88.25	-	67.68	74.16	79.11	84.63
HRNet18 [49]	2019	1.59	78.11	86.81	90.88	96.74	1.84	69.89	78.95	85.16	92.32
Payer <i>et al.</i> [40]	2019	1.34	81.47	89.36	93.15	97.01	1.65	69.94	78.84	85.74	93.89
Chen <i>et al.</i> [10]	2019	1.17	86.67	92.67	95.54	98.53	1.48	75.05	82.84	88.53	95.05
Ours	-	1.04	88.49	93.12	95.72	98.42	1.43	76.57	83.68	88.21	94.31

most of the other state-of-the-art methods in both datasets. In particular for the COFW-68 dataset, the Mean Error and FR@0.1 are significantly improved (5% and 86%) comparing to the previous best model, STYLE [42], demonstrating a strong cross-dataset generalizability of our method.

Cephalometric X-rays: We further applied our model on a public Cephalometric X-ray dataset and compare with HRNet18 [49] and three domain specific state-of-the-art models on this dataset, Arik *et al.* [2], Payer et al. [40] and Chen *et al.* [10]. As is shown in Table 5, our model significantly outperforms Arik *et al.*, HRNet18 [49] and Payer et al. [40] in all metrics. Comparing to Chen *et al.* [10], we also achieve improved overall accuracy evaluated under MRE. A closer look at the error distribution reveals that our model is able to achieve more precise localization under smaller error ranges, i.e., 2mm and 2.5mm.

Hand and Pelvic X-rays: As shown in Table 4, our model achieves substantial performance improvements comparing to the HRNet18 [49], Payer et al. [40] and Chen *et al.* [10] on both the Hand and Pelvic X-ray datasets. On Hand X-ray, where the bone structure can vary in different shapes depending on the hand pose, our method still achieves largely reduced Hausdorff distance as well as its standard deviation, revealing DAG’s ability in capturing landmark relationships under various situations toward robust landmark detection.

4.4 Graph Structure Visualization

To better understand learning outcomes, we look into the visualization on the learned graph structure. As shown in Figure 3, the learned structures in different domains are meaningful indicating strong connections between 1) spatially close landmarks, and 2) remote but related landmarks that move coherently, e.g. symmetrical body parts. We believe the mechanism behind our algorithm is relying on these locations to provide reliable inductions when it makes movement predictions, such as similar movements by neighbors, or fixed spatial relationships by the symmetrical body parts (e.g., eyes, pelvis). With the learnable graph connectivity, we are able to capture the underlying landmarks relationships for different objects.

4.5 Ablation Studies

In this section, we examine the performance of the proposed methods by conducting ablation studies on the 300W fullset. We analyze: 1) the overall effect

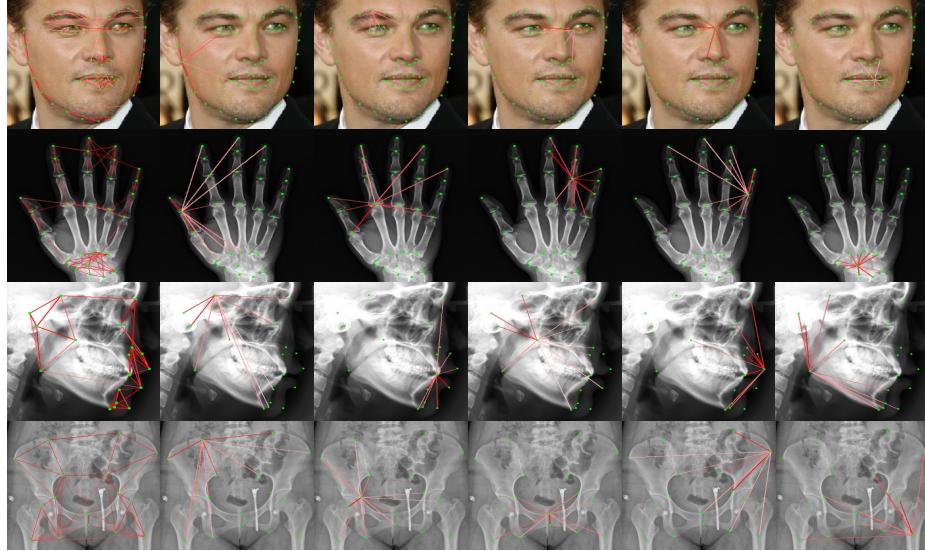


Fig. 3: Graph structure visualization. Red lines: edges. Green dots: landmarks. Deeper red means higher edge weights. [Leftmost column]: the constructed graphs (3 highest weighted edges for each landmark). [Right 5 columns]: for the 5 landmarks, the most related neighbors (10 highest weighted edges).

of using the proposed DAG to regress landmark coordinates, 2) the individual effect of learning the graph connectivity, 3) the individual effect of incorporating shape feature into the graph signal. More ablation studies can be found in the supplementary material.

Overall effect of the proposed DAG: We analyze the effect of using DAG to regress landmark coordinates in comparison with two baselines, namely 1) *Global feature*: The last feature map of the backbone network is global average pooled to produce a feature vector, which connects to a fully connected layer to regress landmark coordinates. This approach is similar to previous coordinate regression based methods, e.g. [63,71]. 2) *Local feature*: The feature vectors are interpolated at each landmark’s initial location on the last feature map of the backbone CNN. Then each landmark’s feature vector is connected to a fully connected layer to regress the landmark’s coordinate. To decouple the effect of the backbone strength, each experiment is conducted on four popular landmark detection backbone networks, namely VGG16 [47], ResNet50 [24], StackedHourGlass4 [38], HRNet18 [49]. Results are listed in Table 6. By comparing different regression methods with the same backbone (columnwise), DAG achieves the best results indicating the proposed framework’s strong localization ability. By comparing DAG’s results under different backbones (last row), we observe DAG’s consistent performance boost demonstrating its effectiveness and promising generalizability.

Table 6: Ablation studies on the effectiveness of the proposed method DAG.

	VGG16	ResNet50	StackedHG4	HRNet18
Global feature	4.66	4.33	4.31	4.30
Local feature	4.42	4.10	3.96	3.72
Proposed DAG	3.66	3.65	3.07	3.04

Table 7: Ablation study on graph connectivity and shape feature.

	w.o Shape Feature	w. Shape Feature
Self	3.31	3.16
Uniform	3.16	3.12
Learned	3.08	3.04

Individual effect of learning graph connectivity: We study three kinds of graph connectivity schemes, namely 1) *Self*-connectivity: The landmarks only connect to themselves and no other landmarks. 2) *Uniform* connectivity: The landmarks connects to all other landmarks using the same edge weight. 3) *Learned* connectivity: learned edge weights as proposed. As summarized in Table 7, regardless of using shape feature or not, using uniform connectivity performs results in better performance than self-connectivity, demonstrating the importance of allowing information exchange on the graph. The learned connectivity performance the best, further demonstrating that learned edge weights further improve the effectiveness of information exchange on the graph. **Individual effect of incorporating shape feature:** We analyze the effect of incorporating the shape feature using self, uniformed and learned connectivities, respectively. As shown in Table 7, on all three types of connectivities, incorporating the proposed shape feature into graph signal results in improved performance especially for self-connective graphs, where the shape feature adds the missing global structure information.

5 Conclusion

In this paper, we introduce a robust and accurate landmark detection model named Deep Adaptive Graph (DAG). The proposed model deploys an initial landmark graph, and then deforms and progressively updates the graph by learning the adjacency matrix. Graph convolution operations follow the strong structural prior to enable effective local information exchange as well as global structural constraints for each step’s movements. The superior performances on three public facial image datasets and three X-ray datasets prove both the effectiveness and generalizability of the proposed method in multiple domains.

Acknowledgement. This work is supported in part by NSF through award IIS-1722847, NIH through the Morris K. Udall Center of Excellence in Parkinson’s Disease Research. The main work was done when Weijian Li was a research intern at PAII Inc.

6 Additional discussions with related works:

Though some recent works [52,8,40] propose to model landmark relationship, our problem/method has large differences from them. Tompson *et al.* [52] propose to use spatial information in a post-processing step to filter outliers, while we leverage visual-spatial joint features for landmark regression. Also, the PAF proposed by Cao *et al.* [8] focuses on a different task of assembling detected key points for multi-person parsing. Zhao *et al.* [72] focus differently on predicting 3D poses from 2D joints. Their 2D joints are generated by a pre-trained 2D pose estimation network. Besides, their network structure is predefined by a fixed adjacency matrix while we actively learn the structures. Payer *et al.* [40], propose a spatial configuration branch to disambiguate candidates from the heatmap predictions. There is no explicit landmark structure modeling. In contrast, we explicitly model shape through a graph representation with learnable connectivity.

Among the SOTA, WING [21] is pure coordinate-based, while LAB [62] and AWING [58] integrate face boundary information via heatmap, which is their key contributions. The gap between WING and AWING is significant on WFLW, which is a more challenging dataset than 300W in terms of dataset scale, pose variations, occlusions, etc. Our method performs significantly better than WING on WFLW by reducing the failure rate by 50%, and is competitive to AWING. In addition, WING focuses on loss design, which is orthogonal and complementary to our novelty. By employing WING loss in our method, our performance can be further improved (e.g., on 300W, inter-pupil NME from 4.27 to 4.21 and inter-ocular NME from 3.04 to 3.01). While LAB and AWING utilize global representation, human knowledge on face structure via a boundary heatmap is injected, leading to task-specific solutions. In contrast, our method is a general landmark detection method to model the structural information via a self-learned graph structure.

7 CED Curve:

Following previous works [62,42], we report Cumulative Errors Distribution (CED) curve result on cross-evaluations of COFW-68 test set. Recall that the success rate measures the proportion of images that have a localization error below a certain threshold [22]. Thus, given a range of thresholds, the corresponding success rates will form a distribution which is considered as Cumulative Error Distribution (CED). For clearer comparison, we include both Normalized Mean Error (**Error**) as well as the Failure Rate (i.e. $1 - SuccessRate$) (**Failure**) at threshold of 0.1. As we can see from Figure 4, our model outperforms previous methods by a large margin, especially in Failure Rate which is reduced to 0.39% for the first time. The comparison of numerical NME and Failure Rate values with the other state-of-the-arts can be found in Table 3 in our submitted ECCV-20 main paper.

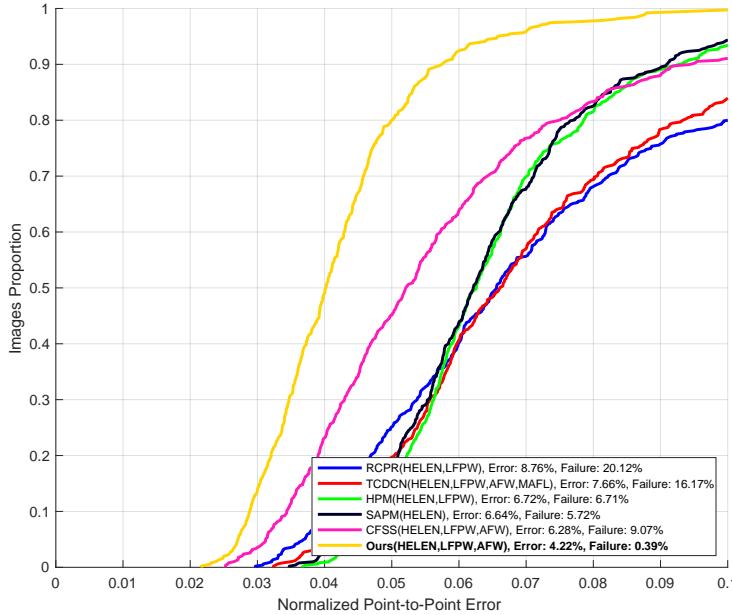


Fig. 4: Cumulative Errors Distribution (CED) curve results on the COFW-68 test set.

8 Ablation Studies

Here we conduct three more types of ablation studies, namely: (1) The comparison of the transformation method used in GCN-global. (2) The effectiveness of the proposed GCN modules.(3) The comparison of different number of regression steps used in GCN-local. Results are recorded in Table 8.

Choice of transformations: We experiment two types of GCN-global choices: (1) Adopt Affine Transformation. In this case, the performance of our GCN-global module drops to 3.13.(2) Adopt Perspective Transformation. We achieve the best result as 3.04 which is also reported in our main paper. This indicates that GCN-global can better locates ROIs with the more flexible perspective transformation.

Effectiveness of GCN modules: We examine the effectiveness of the proposed GCN modules by: (1) Replacing GCN-global with a CNN block; we replace the GCN-global module with a 2-layer CNN (Conv/BN/ReLU) with Global Average Pooling predicting 9 transformation parameters. The average error increased from 3.04 to 3.12. (2) Replacing GCN-local with a MLP block: we remove the connectivity used in GCN-local, making it a simple MLP (FC/ReLU). The average error increased from 3.04 to 3.18. These indicating the importance of the proposed GCN modules.

Table 8: Ablation studies on the proposed model with 300W fullset under Inter-Ocular normalization.

Different Transformations NME	Affine Transformation 3.13	Perspective Transformation (Ours) 3.04		
Effectiveness of GCN modules NME	Replace GCN-global with CNN 3.12	Replace GCN-local with MLP 3.18		
Different GCN Steps NME	Step=1 3.24	Step=3 (Ours) 3.04	Step=5 3.07	Step=7 3.11

Number of steps: We analyze different choices of steps for GCN-local. Results are shown in Table 8. The overall performance improves as the number of steps increases indicating the benefit of cascading multiple regressions. The best performance is achieved when GCN-local is implemented with three iterations.

9 More Settings:

We describe more settings for training the model. Adam optimizer is adopted with initial learning rate $lr = 0.0001$. The learning rate decreases at every 100 epochs. $L2$ penalty is applied to the training parameters with rate 0.0001. Margin for training GCN-global is set to $m = 0.1$ for Face300W, $m = 0.15$ for WFLW, $m = 0.15$ for three Medical datasets. All data augmentations we used: (1) Rotate input image with a random angle in $[-30, 30]$. (2) Random flip the input image horizontally. (3) Scale input image with a random factor in $[0.75, 1.25]$.

References

1. Alp Guler, R., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S., Kokkinos, I.: Densereg: Fully convolutional dense shape regression in-the-wild. In: CVPR. pp. 6799–6808 (2017)
2. Arik, S.Ö., Ibragimov, B., Xing, L.: Fully automated quantitative cephalometry using convolutional neural networks. *Journal of Medical Imaging* **4**(1), 014501 (2017)
3. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In: CVPR. pp. 3444–3451 (2013)
4. Bulat, A., Tzimiropoulos, G.: Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In: ICCV. pp. 3706–3714 (2017)
5. Bulat, A., Tzimiropoulos, G.: Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In: CVPR. pp. 109–117 (2018)
6. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: CVPR. pp. 1513–1520 (2013)
7. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. *IJCV* **107**(2), 177–190 (2014)
8. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. pp. 7291–7299 (2017)
9. Chen, H., Wang, Y., Zheng, K., Li, W., Cheng, C.T., Harrison, A.P., Xiao, J., Hager, G.D., Lu, L., Liao, C.H., et al.: Anatomy-aware siamese network: Exploiting semantic asymmetry for accurate pelvic fracture detection in x-ray images. arXiv preprint arXiv:2007.01464 (2020)
10. Chen, R., Ma, Y., Chen, N., Lee, D., Wang, W.: Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting. In: MICCAI. pp. 873–881. Springer (2019)
11. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. *TPAMI* (6), 681–685 (2001)
12. Cootes, T.F., Taylor, C.J.: Active shape modelssmart snakes. In: BMVC, pp. 266–275. Springer (1992)
13. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. *Computer vision and image understanding* **61**(1), 38–59 (1995)
14. Cristinacce, D., Cootes, T.F.: Feature detection and tracking with constrained local models. In: BMVC. vol. 1, p. 3. Citeseer (2006)
15. Deng, J., Liu, Q., Yang, J., Tao, D.: M3 csr: Multi-view, multi-scale and multi-component cascade shape regression. *Image and Vision Computing* **47**, 19–26 (2016)
16. Deng, J., Trigeorgis, G., Zhou, Y., Zafeiriou, S.: Joint multi-view face alignment in the wild. *TIP* **28**(7), 3636–3648 (2019)
17. DeTone, D., Malisiewicz, T., Rabinovich, A.: Deep image homography estimation. arXiv preprint arXiv:1606.03798 (2016)
18. Dong, X., Yan, Y., Ouyang, W., Yang, Y.: Style aggregated network for facial landmark detection. In: CVPR. pp. 379–388 (2018)
19. Dong, X., Yu, S.I., Weng, X., Wei, S.E., Yang, Y., Sheikh, Y.: Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In: CVPR. pp. 360–368 (2018)

20. Fan, H., Zhou, E.: Approaching human level facial landmark localization by deep learning. *Image and Vision Computing* **47**, 27–35 (2016)
21. Feng, Z.H., Kittler, J., Awais, M., Huber, P., Wu, X.J.: Wing loss for robust facial landmark localisation with convolutional neural networks. In: CVPR. pp. 2235–2245 (2018)
22. Ghiasi, G., Fowlkes, C.C.: Occlusion coherence: Detecting and localizing occluded faces. arXiv preprint arXiv:1506.08347 (2015)
23. Han, D., Gao, Y., Wu, G., Yap, P.T., Shen, D.: Robust anatomical landmark detection with application to mr brain image registration. *Computerized Medical Imaging and Graphics* **46**, 277–290 (2015)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
25. Honari, S., Molchanov, P., Tyree, S., Vincent, P., Pal, C., Kautz, J.: Improving landmark localization with semi-supervised learning. In: CVPR. pp. 1546–1555 (2018)
26. Honari, S., Yosinski, J., Vincent, P., Pal, C.: Recombinator networks: Learning coarse-to-fine feature aggregation. In: CVPR. pp. 5743–5752 (2016)
27. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: NeurIPS. pp. 2017–2025 (2015)
28. Kumar, A., Marks, T.K., Mou, W., Wang, Y., Jones, M., Cherian, A., Koike-Akino, T., Liu, X., Feng, C.: Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In: CVPR. pp. 8236–8246 (2020)
29. Kumar, A., Chellappa, R.: Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. In: CVPR. pp. 430–439 (2018)
30. Li, G., Müller, M., Thabet, A., Ghanem, B.: Can gcn go as deep as cnns? CVPR (2019)
31. Lindner, C., Bromiley, P.A., Ionita, M.C., Cootes, T.F.: Robust and accurate shape model matching using random forest regression-voting. *TPAMI* **37**(9), 1862–1874 (2014)
32. Ling, H., Gao, J., Kar, A., Chen, W., Fidler, S.: Fast interactive object annotation with curve-gcn. In: CVPR. pp. 5257–5266 (2019)
33. Liu, X.: Generic face alignment using boosted appearance model. In: CVPR. pp. 1–8. IEEE (2007)
34. Liu, Z., Yan, S., Luo, P., Wang, X., Tang, X.: Fashion landmark detection in the wild. In: ECCV. pp. 229–245. Springer (2016)
35. Lu, Y., Zheng, K., Li, W., Wang, Y., Harrison, A.P., Lin, C., Wang, S., Xiao, J., Lu, L., Kuo, C.F., et al.: Learning to segment anatomical structures accurately from one exemplar. arXiv preprint arXiv:2007.03052 (2020)
36. Lv, J., Shao, X., Xing, J., Cheng, C., Zhou, X.: A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In: CVPR. pp. 3317–3326 (2017)
37. Milborrow, S., Nicolls, F.: Locating facial features with an extended active shape model. In: ECCV. pp. 504–513. Springer (2008)
38. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV. pp. 483–499. Springer (2016)
39. Payer, C., Štern, D., Bischof, H., Urschler, M.: Regressing heatmaps for multiple landmark localization using cnns. In: MICCAI. pp. 230–238. Springer (2016)
40. Payer, C., Štern, D., Bischof, H., Urschler, M.: Integrating spatial configuration into heatmap regression based CNNs for landmark localization. *MIA* **54**, 207–219 (may 2019). <https://doi.org/10.1016/j.media.2019.03.007>

41. Qi, M., Li, W., Yang, Z., Wang, Y., Luo, J.: Attentive relational networks for mapping images to scene graphs. In: CVPR. pp. 3957–3966 (2019)
42. Qian, S., Sun, K., Wu, W., Qian, C., Jia, J.: Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In: ICCV. pp. 10153–10163 (2019)
43. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features. In: CVPR. pp. 1685–1692 (2014)
44. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: CVPRW. pp. 397–403 (2013)
45. Saragih, J.M., Lucey, S., Cohn, J.F.: Face alignment through subspace constrained mean-shifts. In: ICCV. pp. 1034–1041. IEEE (2009)
46. Sauer, P., Cootes, T.F., Taylor, C.J.: Accurate regression procedures for active appearance models. In: BMVC. pp. 1–11 (2011)
47. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
48. Su, J., Wang, Z., Liao, C., Ling, H.: Efficient and accurate face alignment by global regression and cascaded local refinement. In: CVPRW (2019)
49. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR. pp. 5693–5703 (2019)
50. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: CVPR. pp. 3476–3483 (2013)
51. Tang, Z., Peng, X., Geng, S., Wu, L., Zhang, S., Metaxas, D.: Quantized densely connected u-nets for efficient landmark localization. In: ECCV. pp. 339–354 (2018)
52. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: NeurIPS. pp. 1799–1807 (2014)
53. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: CVPR. pp. 1653–1660 (2014)
54. Trigeorgis, G., Snape, P., Nicolaou, M.A., Antonakos, E., Zafeiriou, S.: Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In: CVPR. pp. 4177–4187 (2016)
55. Valle, R., Buenaposada, J.M., Valdés, A., Baumela, L.: A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment. In: ECCV. pp. 585–601 (2018)
56. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
57. Wang, C.W., Huang, C.T., Lee, J.H., Li, C.H., Chang, S.W., Siao, M.J., Lai, T.M., Ibragimov, B., Vrtovec, T., Ronneberger, O., et al.: A benchmark for comparison of dental radiography analysis algorithms. MIA **31**, 63–76 (2016)
58. Wang, X., Bo, L., Fuxin, L.: Adaptive wing loss for robust face alignment via heatmap regression. In: ICCV. pp. 6971–6981 (2019)
59. Wang, Y., Lu, L., Cheng, C.T., Jin, D., Harrison, A.P., Xiao, J., Liao, C.H., Miao, S.: Weakly supervised universal fracture detection in pelvic x-rays. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 459–467. Springer (2019)
60. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR. pp. 4724–4732 (2016)
61. Wu, S., Tang, Y., Zhu, Y., Wang, L., Xie, X., Tan, T.: Session-based recommendation with graph neural networks. In: AAAI. vol. 33, pp. 346–353 (2019)

62. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: A boundary-aware face alignment algorithm. In: CVPR. pp. 2129–2138 (2018)
63. Wu, W., Yang, S.: Leveraging intra and inter-dataset variations for robust face alignment. In: CVPRW. pp. 150–159 (2017)
64. Wu, Y., Ji, Q.: Facial landmark detection: A literature survey. IJCV **127**(2), 115–142 (2019)
65. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: CVPR. pp. 532–539 (2013)
66. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? arXiv preprint arXiv:1810.00826 (2018)
67. Yu, W., Liang, X., Gong, K., Jiang, C., Xiao, N., Lin, L.: Layout-graph reasoning for fashion landmark detection. In: CVPR. pp. 2937–2945 (2019)
68. Yu, X., Huang, J., Zhang, S., Metaxas, D.N.: Face landmark fitting via optimized part mixtures and cascaded deformable model. TPAMI **38**(11), 2212–2226 (2015)
69. Yu, X., Zhou, F., Chandraker, M.: Deep deformation network for object landmark localization. In: ECCV. pp. 52–70. Springer (2016)
70. Zhang, J., Shan, S., Kan, M., Chen, X.: Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In: ECCV. pp. 1–16. Springer (2014)
71. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Learning deep representation for face alignment with auxiliary attributes. TPAMI **38**(5), 918–930 (2015)
72. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3d human pose regression. In: CVPR. pp. 3425–3435 (2019)
73. Zhou, F., Brandt, J., Lin, Z.: Exemplar-based graph matching for robust facial landmark localization. In: ICCV. pp. 1025–1032 (2013)
74. Zhu, M., Shi, D., Zheng, M., Sadiq, M.: Robust facial landmark detection via occlusion-adaptive deep networks. In: CVPR. pp. 3486–3496 (2019)
75. Zhu, S., Li, C., Change Loy, C., Tang, X.: Face alignment by coarse-to-fine shape searching. In: CVPR. pp. 4998–5006 (2015)
76. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: CVPR. pp. 146–155 (2016)
77. Zhu, Z., Luo, P., Wang, X., Tang, X.: Deep learning identity-preserving face space. In: ICCV. pp. 113–120 (2013)
78. Zou, X., Zhong, S., Yan, L., Zhao, X., Zhou, J., Wu, Y.: Learning robust facial landmark detection via hierarchical structured ensemble. In: ICCV (2019)