Robust Facial Landmark Detection via Aggregation on Geometrically Manipulated Faces

Seyed Mehdi Iranmanesh, Ali Dabouei, Sobhan Soleymani, Hadi Kazemi, Nasser M. Nasrabadi
West Virginia University
{seiranmanesh,ad0046,ssoleyma,hakazemi}@mix.wvu.edu, {nasser.nasrabadi}@mail.wvu.edu

Abstract

In this work, we present a practical approach to the problem of facial landmark detection. The proposed method can deal with large shape and appearance variations under the rich shape deformation. To handle the shape variations we equip our method with the aggregation of manipulated face images. The proposed framework generates different manipulated faces using only one given face image. The approach utilizes the fact that small but carefully crafted geometric manipulation in the input domain can fool deep face recognition models. We propose three different approaches to generate manipulated faces in which two of them perform the manipulations via adversarial attacks and the other one uses known transformations. Aggregating the manipulated faces provides a more robust landmark detection approach which is able to capture more important deformations and variations of the face shapes. Our approach is demonstrated its superiority compared to the state-of-the-art method on benchmark datasets AFLW, 300-W, and COFW.

1. Introduction

Facial landmark detection goal is to identify the location of predefined facial landmarks (*i.e.*, tip of the nose, corner of the eyes, and eyebrows). Reliable landmark estimation is part of the procedure for more complicated vision tasks. It can be applied to the variant tasks such as 3D face reconstruction [30], head pose estimation [48], facial reenactment [44], and face recognition [61]. However, it remains challenging due to the necessity of handling non-rigid shape deformations, occlusions, and appearance variations. For example, facial landmark detection must handle not only coarse variations such as illumination and head pose but also finer variations including skin tones and expressions. There has been a wide range of approaches to solve the problem of landmark detection, starting with methods such as active shape model [8], and active appearance models [7]

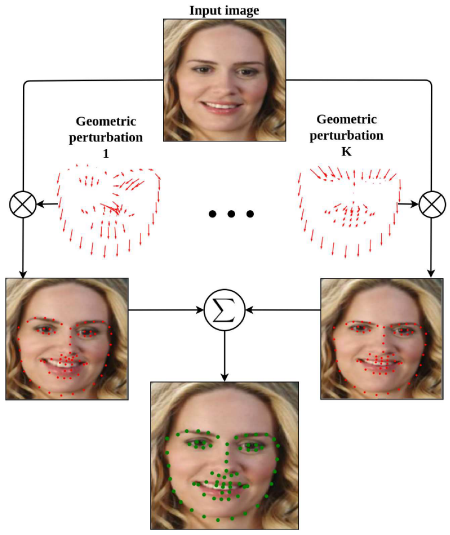

Figure 1: An input face image is manipulated utilizing geometric perturbations that target important locations of face images for the task of landmark detection. K different manipulated faces are generated where each of them contains the important displacements from the input image. The aggregation on these manipulated images leads to robust landmark detection.

which are related to PCA-based shape constraint.

Many of these approaches utilize a cascade strategy to integrate prediction modules and update the landmark locations in a progressive manner [54, 11]. Cascade regression networks which are designed for landmark localization [42], or human body pose estimation [45] have made improvements by tackling the problem level at coarse to fine levels. However, requiring careful design and initialization for such frameworks and the absence of learned geometric relationships are the main challenges of these architectures.

Recently, with the onset of convolutional neural networks (ConvNets) in feature representation [40], a common approach in facial landmark detection is to extract features

978-1-7281-6553-0/20/$31.00 ©2020 IEEE

from the facial appearance using ConvNets, and afterward learn a model typically a regressor to map the features to the landmark locations [54, 10, 33, 36]. Despite the excellent performance of the ConvNets in different applications, it has been shown [17, 43] that they can be very sensitive and vulnerable to a small perturbation in the input domain which can lead to a drastic change of the output domain, *e.g.*, predicted landmarks.

Many approaches solve the face alignment problem with multi-tasking approaches. However, the task of face alignment might not be in parallel with the other tasks. For example, in the classification task, the output needs to be invariant to small deformations such as translation. However, in tasks such as landmark localization or image segmentation both the global integration of information as well as maintaining the local information and pixel-level detail is necessary. The goal of precise landmark localization has led to evolving new architectures such as dilated convolutions [52], recombinator-networks [21], stacked what where auto-encoders [58], and hyper-columns [18] where each of them attempts to preserve pixel-level information.

In this paper, we propose a geometry aggregated network (GEAN) for face alignment which can comfortably deal with rich expressions and arbitrary shape variations. We design a novel aggregation framework which optimizes the landmark locations directly using only one image without requiring any extra prior which leads to robust alignment given arbitrary face deformations. We provide three different approaches to produce deformed images using only one image and aggregate them in a weighted manner according to their amount of displacement to estimate the final locations of the landmarks. Extensive empirical results indicate the superiority of the proposed method compared to existing methods on challenging datasets with large shape and appearance variations, *i.e.*, 300-W [38] and ALFW [27].

2. Related Work

A common approach to facial landmark detection problem is to leverage deep features from ConvNets. These facial features and regressors are trained in an end-to-end manner utilizing a cascade strategy to update the landmark locations progressively [42, 60]. Yu et al. [53] integrate geometric constraints within CNN architecture using a deep deformation network. Lev et al. [31] propose a deep regression framework with two-step re-initialization to avoid the initialization issue. Zhu et al. [1] also tried to deal with poor initialization utilizing a coarse search over a shape space with variant shapes. In another work, Zhu et al. [60], overcome the extreme head poses and rich shape deformations exploiting cascaded regressors.

Another category of landmark detection approaches leverages the end-to-end training from ConvNets frameworks to learn robust heatmaps for landmark detection task [47, 4, 32]. Balut et al. [4] utilized the residual framework to propose a robust network for facial landmark detection. Newell et al. [32] and Wei et al. [47] consider the coordinate of the highest response on the heatmaps as the location of landmarks for human pose estimation task.

In a more general definition, this problem can also be viewed as learning structural representation. Some studies [34, 35], disentangle visual content into different factors of variations such as camera viewpoint, motion and identity to capture the inherent structure of objects. However, the physical parameters of these factors are embedded in a latent representation which is not discernible. Some methods can handle [55, 19] conceptualize structures in the multi-tasking framework as auxiliary information (*e.g.*, landmarks, depth, and mask). Such structures in these frameworks are designed by humans and need supervision to learn.

3. Proposed Method

Given a face image $I \in \mathbb{R}^{w \times h}$ with spatial size $W \times H$, the facial landmark detection algorithm aims to find a prediction function $\Phi : \mathbb{R}^{W \times H} \rightarrow \mathbb{R}^{2 \times L}$ which estimates the 2D locations of L landmarks. We seek to find a robust and accurate version of Φ by training a deep function through the aggregation of geometrically manipulated faces. The proposed method consists of different parts which will be described in detail.

3.1. Aggregated Landmark Detector.

The proposed approach attempts to provide a robust landmark detection algorithm to compensate for the lack of a specific mechanism to handle arbitrary shape variations in the literature of landmark detection. The method builds upon aggregating set of manipulated images to capture robust landmark representation. Given a face image I, a set of manipulated images are constructed such that $\hat{I}_k = M(I, \theta_k)$ is the k-th manipulated face image and θ_k is its related parameters for the manipulating function M. Considering the set of manipulated images, we seek a proper choice of M such that aggregating landmark information in the set $\{\Phi(\hat{I}) : k = 1 \dots K\}$ provides a more accurate and robust landmark features compared to $\Phi(I)$ which solely uses the original image I. Therefore, one important key in the aggregated method is answering the question of "how" to manipulate images. Face images typically have a semantic structure which have a similar global structure but the local and relative characteristics of facial regions differ between individuals. Hence, a straightforward and comprehensive choice of the manipulation function M should incorporate the prior information provided by the global consistency of semantic regions and uniqueness of relative features which can be interpreted as the ID information. Hence, we build our work based on a choice of M

which incorporates geometric transformations to manipulate relative characteristics of inputs samples while preserving the semantic and global structure of input faces.

To incorporate ID information, we consider a pretrained face recognizer $f : \mathbb{R}^{W \times H} \rightarrow \mathbb{R}^{n_z}$ mapping an input face image to an ID representation $z \in \mathbb{R}^{n_z}$, where cardinality of the embedding subspace is n_z (typically set to be 128 [39]). Having f makes it possible to compare IDs of two samples by simply measuring the ℓ_2-norm of their representation in the embedding space. Hence, we geometrically manipulate the input face image to change its ID. It should be noted that since f is trained on face images, the corresponding embedding space of IDs captures a meaningful representation of faces. Therefore, the manipulated faces contain rich information with regards to face IDs.

To manipulate the face image I based on landmark coordinates, we consider coarse landmark locations $P = \{(x_0, y_0), \ldots, (x_{L-1}, y_{L-1})\}$ and define the displacement field d to manipulate the landmark locations. Given the i-th source landmark (x_i, y_i), we compute its manipulated version using the displacement vector $d_i = (\Delta x_i, \Delta y_i)$. The manipulated landmark $p_i + d_i$ is as follows:

$$p_i + d_i = (x_i + \Delta x_i, y_i + \Delta y_i) \ . \tag{1}$$

We present three different approaches to find a proper displacement (d) for manipulating face images.

3.2. Manipulation by Adversarial Attack.

In the first approach we use adversarial attacks [16] to manipulate facial landmarks to fool a face recognizer. Xiao et al. [49], proposed stAdv attack to generate adversarial examples using spatially transforming benign images. They utilize a displacement field for all the pixels in the input image. Afterward, they computed the corresponding location of pixels in the adversarial image using the displacement field d. However, optimizing a displacement field for all the pixels in the image is a highly non-convex function. Therefore, they used the L-BFGS [29], with a linear backtrack search to find the optimal displacement field which is computationally expensive. Here, our approach considers the fact that the facial landmarks provide highly discriminative information for face recognition tasks [23]. In fact, face recognition tasks are highly linear around the original coordinates of the facial landmarks as it is shown in [9].

In contrast to [49] which computes the displacement field for all the pixels, our proposed method is inspired by [9] and estimates the d only for L landmarks and it does not suffer from the computational complexity. In addition, it is possible to apply the conventional spatial transformation to transform image. Therefore, the adversarial (manipulated) image using the transformation T is as follows:

$$\hat{I} = T(P, P + d, I) \ , \tag{2}$$

where T is the thin plate spline (TPS) [3] transformation mapping from the source landmarks (control points) P to the target ones $P + d$. In order to make the whole framework differentiable with respect to the landmark locations, we select a differentiable interpolation function (*i.e.*, differentiable bilinear interpolation) [24] so that the prediction of the face recognizer is differentiable with respect to the landmark locations.

In this approach, we employ the gradient of the prediction in a face recognition model to update the displacement field d and geometrically manipulate the input face image. We extend Dabouei et al. [9] work in a way to generate K different adversarial faces where each face represents a different ID (K different IDs will be generated). Considering an input image I, a face recognizer f , and a set of $k-1$ manipulated images $S_I = \{\hat{I}_1, ..., \hat{I}_{k-1}\}$ the cost is defined as follows for the k-th adversarial face:

$$\mathcal{L} = \sum_{I' \in S_I} ||f(T(P, P + d, I)) - f(I')||_2 \ . \tag{3}$$

Inspired by FGSM [16], we employ the direction of the gradients of the prediction to update the adversarial landmark locations $P + d$, in an iterative manner. Considering $P + d$ as P^{adv}, using FGSM [16], the t-th step of optimization is as follows:

$$P_t^{adv} = P_{t-1}^{adv} + \epsilon \ sign(\nabla_{P_{t-1}^{adv}} \mathcal{L}) \ . \tag{4}$$

In addition, we consider the clipping technique to constrain the displacement field in order to prevent the model from generating distorted face images. The algorithm continues the optimization for the k-th landmark locations until $\min_{I' \in S_I} \{||f(\hat{I}) - f(I')||_2\} < \tau$ is failed, where τ is simply the distance threshold in the embedding space. In this way, we make sure that the k-th manipulated face has a minimum distance of τ to the other manipulated images in the face embedding subspace. Algorithm 1 shows the proposed procedure for generating K different manipulated faces.

3.3. Manipulation of Semantic Groups of Landmarks using Adversarial Attacks.

In the first approach, we consider a fast and efficient approach to generate different faces based on the given face image. However, the first approach does not directly consider the fact that different landmarks semantically placed in different groups (*i.e.*, landmarks related to lip, left eye, right eye, etc.). This might lead to generating severely distorted adversarial images.

We added the clipping constraint to mitigate this issue in the first approach. Here, we perform semantic landmarks grouping [9]. We categorize the landmarks into n semantic groups $P_i, i \in \{1, \ldots n\}$, where $p_{i,j}$ denotes the j-th landmark in the group i which contains c_i landmarks.

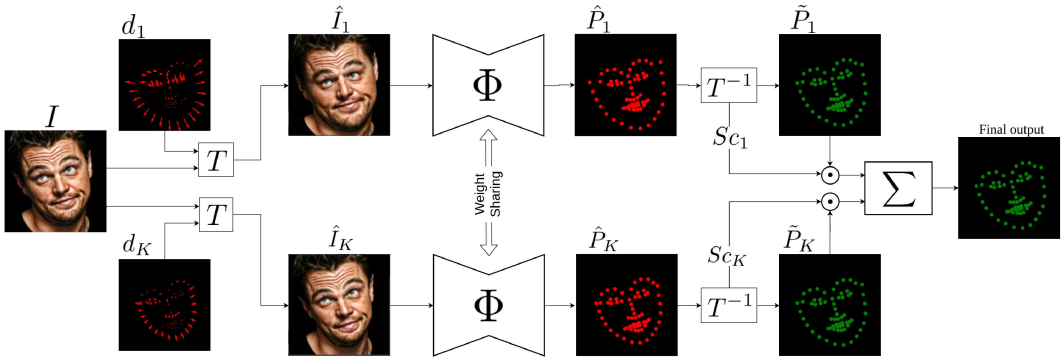

Figure 2: Overview of the proposed aggregated framework (GEAN). It consists of four steps: 1) K different manipulated faces are generated; 2) Each manipulated face is given to the shared landmark detector Φ to extract its landmarks; 3) The inverse of transformation matrix is applied to the extracted landmarks to compensate for the displacement of step 1; 4) The normalization score values for each landmark of each branch is calculated and the aggregation is performed to extract the final landmark locations.

Algorithm 1: Adversarial Face Generation

1 **Input:** Image I, number of branches K, face recognizer f, distance threshold τ, clipping threshold δ.
2 **Output:** Set of adversarial faces $S = \{\hat{I}_1, ..., \hat{I}_K\}$.
3 Initialize $\hat{I} \leftarrow I$ and $S = \{I\}$.
4 **for** $k = 1$ **to** K **do**
5 $\quad \hat{I}_{t=0,k} \leftarrow I$;
6 $\quad$ **while** $\min_{I' \in S_I} \{||f(\hat{I}_{t,k}) - f(I')||_2\} < \tau$ **do**
7 $\quad\quad \mathcal{L} = \sum_{I' \in S_I} ||f(T(P, P_t^{adv}, I)) - f(I')||_2$;
8 $\quad\quad P_{t+1}^{adv} = P_t^{adv} + \epsilon \; sign(\nabla_{P_t^{adv}} \mathcal{L})$;
9 $\quad\quad P_{t+1}^{adv} = clip\ (P_{t+1}^{adv}, \delta)$;
10 $\quad\quad \hat{I}_{t+1,k} = T(P, P_{t+1}^{adv}, I)$;
11 $\quad S \leftarrow \{S, \hat{I}_k\}$;
12 return $S - \{I\}$;

These groups are formed based on different semantic regions which construct the face shape (*i.e.*, lip, left eye, right eye, etc.). Semantic landmark grouping considers a scale and translation for each semantic group, instead of independently displacing each landmark. This consideration allows us to increase the total amount of displacement while preserving the global structure of the face.

Let P_i represents the i-th landmark group (*e.g.*, group of landmarks related to the lip). The adversarial landmark locations which are semantically grouped can be obtained as following:

$$P_i^{adv} = \alpha_i(P_i - \bar{p_i}) + \beta_i \ , \tag{5}$$

where $\bar{p_i} = \frac{1}{c_i} \sum_{j=1}^{c_i} p_{i,j}$ is the average location of all the landmarks in group P_i, and α_i and β_i for each group can be computed using the closed-from solution in [9]. It should be noted that the value of displacement d for each branch used for computing semantic scales and translations is obtained using Algorithm 1. The only difference is that we add semantic perturbations constructed by Eq. 5 instead of the random perturbation in line 8 of Algorithm 1. Therefore, the scale and translation of each semantic part of the face is different from other manipulated images in the set S.

3.4. Manipulation of Semantic Group of Landmarks with Known Transformation.

In this approach, we semantically group the landmark locations in the same manner as the previous approach. Afterward, we uniformly sample ranges $[0.9, 1.1]^2$ and $[-0.05 \times W, 0.05 \times W]^2$ for the scale and translation of each semantic group, respectively. It may be noted that in the post-processing stage, we make sure that the semantic inter-group structure is preserved, *e.g.*, eyes region does not interfere with the eyebrows region and they are symmetric according to each other. Therefore, the heuristic post-processing limits the above ranges based on the properties of each group. For instance, eyebrows could achieve higher vertical displacement compared to eyes since there is no semantic part above them to impose a constraint.

3.5. Landmark Detector.

Next, the hourglass network proposed in [32] is employed to estimate the facial landmarks location. Hourglass is designed based on residual blocks [20]. It is a symmetric top-down and bottom-up fully convolutional network. The residual modules are able to capture high-level features based on the convolutional operation, while they can maintain the original information with the skip connections. The original information is branched out before downsampling

and concatenated together before each up-sampling to retain the resolution information. Therefore, hourglass is an appropriate topology to capture and consolidate information from different resolutions and scales.

After manipulating the face images (employing either of the three aforementioned approaches), we employ the hourglass network, Φ, to extract the landmarks from the manipulated images. The network Φ is shared among all the branches of the framework as it is shown in Fig. 2. Each landmark has a corresponding detector, which convolutionally extracts a response map. Taking r_i as a i-th response map, we use the weighted mean coordinate as the location of the i-th landmark as follows:

$$\hat{p}_i = (x_i, y_i) = \frac{1}{\zeta_i} \sum_{u=1}^{H} \sum_{v=1}^{W} (u, v).r_i(u, v) \ , \tag{6}$$

where H and W are the height and width of the response map which are the same as the spatial size of the input image, and $\zeta_i = \sum_{u,v} r_i(u, v)$.

3.6. Aggregation.

After extracting the facial landmarks using the shared landmark detector Φ for each of the manipulated face images, we aim to move the predicted landmarks $\hat{P}$ toward their original locations. Let T be a transformation that is used to convert the original faces to the manipulated ones (*i.e.*, via adversarial attack approaches or the known transformation approach). We employ the inverse of the transformation matrix on the predicted landmarks to compensate for the displacement of them and denote the new landmark locations as $\tilde{P}$.

The proposed approach contains a set of landmarks from K branches, *i.e.*, $\tilde{P} = \{\tilde{p}_{i,k}\}$ in which $i \in \{1, \ldots, L\}$, and $k \in \{1, \ldots, K\}$ is the i-th landmark location in k-th branch of the framework. Each branch considers a score value which normalizes the displacement of landmarks caused by the manipulation approach (*i.e.*, via adversarial attacks or known transformations) in each branch of the aggregated network as follows:

$$Sc_{i,k} = \frac{\sqrt{\Delta x_{i,k}^2 + \Delta y_{i,k}^2}}{\sum_{k=1}^{K} \sqrt{\Delta x_{i,k}^2 + \Delta y_{i,k}^2}} \ , \tag{7}$$

where $Sc_{i,k}$ represents the displacement value for the i-landmark in the k-th branch. This score is utilized as a weight to cast appropriate loss punishment in different branches during the optimization of the proposed aggregated landmark detection as follows:

$$\mathcal{L}_T = \frac{1}{LK} \sum_{i=1}^{L} \sum_{k=1}^{K} Sc_{i,k} ||p_i^* - \tilde{p}_{i,k}||_2 \ , \tag{8}$$

where $\tilde{p}_{i,k}$ represents the i-th estimated landmark at k-th branch and p_i^* indicates the ground truth for i-th landmark location.

Given a test image, we extract the rough estimation of the landmark coordinates employing the trained landmark detector Φ in the aggregated approach and consider them as the coarse landmarks P. Afterward, we perform the manipulation approach on the extracted landmarks P and generate manipulated images. The extracted landmarks and manipulated images are used in the aggregated framework to produce the final landmarks. The final landmarks are calculated as follows:

$$p_i^f = \sum_{k=1}^{K} Sc_{i,k}.(\tilde{p}_{i,k}) \ , \tag{9}$$

where p_i^f is the coordinate of the i-th landmark employing the proposed aggregated network such that $\Phi(I) = P^f$ for L landmark locations.

As it is mentioned in the manuscript, during the training phase the manipulation is performed on the coarse landmarks' locations and we have access to them. However, given a test image, we extract the landmarks using the trained landmark detector Φ and then use them as the coarse landmarks' locations in the aggregated framework to predict the final landmark locations.

One question that comes to mind is: what if the predicted landmark locations using the trained landmark detector Φ are not an accurate representation for the original coarse landmarks? To compensate this issue and make the conditions equal for the training and testing phases, we add random noise to the ground truth landmarks such that $P = P^* + \eta$ where P^* is the ground truth for landmark coordinates and η is random noise. Afterward, we employ these landmarks as the coarse landmarks P in the aggregated framework during the training phase.

4. Experiments

In the following section, we consider three variations of our GEAN approach. $GEAN_{adv}$ (3.2) represents the case when the manipulated faces are generated using the adversarial attack approach. $GEAN_{Gadv}$ (3.3) and $GEAN_{GK}$ (3.4) represent the cases when the manipulated faces are generated using the semantically grouped adversarially attack and known transformations approach, respectively. In order to show the effectiveness of GEAN we evaluate its performance on three following datasets:

300-W [38]: The dataset annotates five existing datasets with 68 landmarks: LFPW [2], AFW [62], HELEN [28], iBug, and XM2VTS. Following the common setting in [12, 31], we consider 3,148 training images from LFPW, HELEN, and the full set of AFW. The testing dataset is split into three categories of common, challenging, and full

Figure 3: The representative results for three face images from the 300-W dataset. For each face, the first row represents displacement fields for the aggregated network with $K = 3$ (the arrows are exaggerated for the sake of illustration). The second row shows manipulated images using the corresponding displacement field, and the third row represents the extracted landmarks given the corresponding manipulated images to the landmark detector,$\Phi(\hat{I})$. The fourth row represents landmarks' locations on the input image I from the base detector (in blue), ground-truth (in green), and GEAN landmark detector (in magenta), respectively.

Methods	ERT [26]	LBF [37]	CFSS [1]	CCL [60]	Two-St. [31]	SAN [12]	ODN [59]	LRef. [41]	GEAN_{adv}	GEAN_{GK}	GEAN_{Gadv}
AFLW-Full	4.35	4.25	3.92	2.72	2.17	1.91	1.63	1.63	1.69	1.64	**1.59**
AFLW-Front	2.75	2.74	2.68	2.17	-	1.85	1.38	1.46	1.44	1.38	**1.34**

Table 1: Comparison of different methods based on normalized mean errors (NME) on AFLW dataset.

groups. The common group contains 554 testing images from LFPW and HELEN datasets, and the challenging test set contains 135 images from the IBUG dataset. Combining these two subsets form the full testing set.

AFLW [27]: This dataset contains 21,997 real-world images with 25,993 faces in total with a large variety in appearance (*e.g.*, pose, expression, ethnicity, and age) and environmental conditions. This dataset provides at most 21 landmarks for each face. Having faces with different pose, expression, and occlusion makes this dataset challenging to train a robust detector. Following the same setting as in [31, 12], we do not consider the landmark of two ears. This dataset has two different categories of AFLW-Full and AFLW-Frontal [60]. AFLW-Full contains 20,000 training samples and 4,386 testing samples. AFLW-Front uses the same set of training samples as in AFLW-Full, but only contains 1,165 samples with the frontal face for the testing set.

COFW [5]: This dataset contains 1,345 images for training and 507 images for test. Originally this dataset annotated with 21 landmarks for each face. However, there is a new version of annotation for this dataset with 68 landmarks for each face [14]. We used a new version of annotation to evaluate proposed method and comparison with the other methods.

Evaluation: Normalized mean error (NME) and and Cumulative Error Distribution (CED) curve are usually used as metric to evaluate performance of different methods [60, 31]. Following [37], we use the inter-ocular distance to normalize mean error on 300-W dataset. For the AFLW dataset we employ the face size to normalize mean error as there are many faces with inter-ocular distance closing to zero in this dataset [31].

Implementation Details: We employ the face recognition model developed by Schroff et al. [39] which obtain the state-of-the-art accuracy on the Labeled Faces in the Wild (LFW) [22] dataset as the face recognizer. We train this model on more than 3.3M training images and the average of 360 images per ID (subject) from VGGFace2 dataset [6] to recognize 9,101 celebrities. The landmarks are divided to five different categories based on facial regions as: 1) P_1 : right eye and eyebrow, 2) P_2 : left eye and eyebrow, 3) P_3 : nose, 4) P_4 : mouth, and 5) P_5 : jaw. The number of landmarks in each group is as: $\{n_1 = 11, n_2 = 11, n_3 = 9, n_4 = 20, n_5 = 17\}$. We set $\tau = 0.6$, δ to 5% of the width of the bounding box of each face.

The landmarks' coordinates are scaled to lie inside the range $[-1, 1]^2$ where $(-1, -1)$ is the top left corner and $(1, 1)$ is the bottom right corner of the face image. All the

Method	Common	Challenging	Full Set
LBF [37]	4.95	11.98	6.32
CFSS [1]	4.73	9.98	5.76
MDM [46]	4.83	10.14	5.88
TCDCN [56]	4.80	8.60	5.54
Two-Stage [31]	4.36	7.42	4.96
RDR [50]	5.03	8.95	5.80
Pose-Invariant [25]	5.43	9.88	6.30
SAN [12]	3.34	6.60	3.98
ODN [59]	3.56	6.67	4.17
LRefNets [41]	2.71	4.78	3.12
GEAN	**2.68**	**4.71**	**3.05**

Table 2: Normalized mean errors (NME) on 300-W dataset.

coordinates are assumed to be continuous values since TPS has no restriction on the continuity of the coordinates because of the differentiable bilinear interpolation [24]. The face images are cropped and resized to (256×256). We follow the same setting in [51] and use four stacks of hourglass network for the landmark detection network. We train our model with the batch size of 8, weight decay of 5×10^{-4}, and the starting learning rate of 5×10^{-5} on two GPUs. The face bounding boxes are expanded by the ratio of 0.2 and random cropping is performed as data augmentation.

4.1. Comparison with State-of-the-arts Methods:

Results on 300-W. Table 2 shows the performance of different facial landmark detection methods on 300-W dataset. We compare our method to the most recent state-of-the-art approaches in the literature [12, 50, 25, 41]. The number of branches in training and testing phases is set to $K = 5$. Among the three proposed approaches, we consider $GEAN_{Gadv}$ as the final proposed method to compare with the state-of-the-art methods. The results show its superiority compared to the other methods for both types of bounding boxes. The superiority of the proposed method shows the effect of manipulated images which target the important locations in the input face image. Aggregation of these images improves the facial landmark detection by giving more attention to the keypoint locations of the face images.

Results on AFLW. We conduct our experiments on the training/testing splits and the bounding box provided from [60, 1]. Table 1 shows the effectiveness of proposed GEAN. AFLW dataset provides a comprehensive set of unconstrained images. This dataset contains challenging images with rich facial expression and poses up to $\pm 120^{\circ}$ for yaw and $\pm 90^{\circ}$ for pitch and roll. Evaluation of proposed method on this challenging dataset shows its robustness to large pose variations. Indeed, the weighted aggregation of predictions obtained on the set of deformed faces reduces the sensitivity of GEAN to large pose variations.

Results on COFW. Figure 4 shows the evaluation of our proposed method in a cross-dataset scenario. We conduct

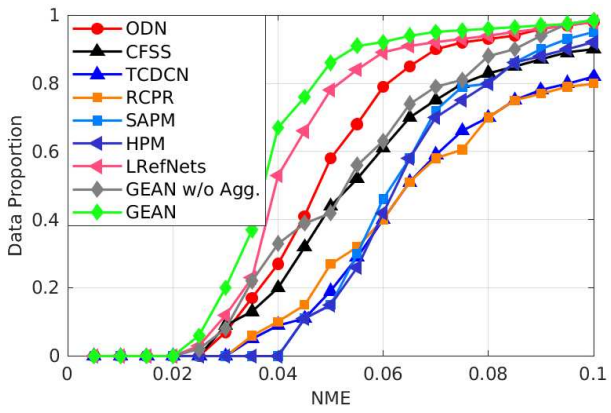

Figure 4: Comparison results of different methods (ODN [59], CFSS [1], TCDCN [57], RCPR [5], SAPM [15], HPM [13], LRefNets [41], and GEAN) on COFW dataset.

evaluation using the models trained on the 300-W dataset and test them on re-annotated COFW dataset with 68 landmarks [14]. The comparison is performed using the CED curves as plotted in Figure 4. The best performance belongs to our method (GEAN) with 4.24% mean error compared to the previous best [41] with 4.40% mean error. This shows the robustness of our method compared to other state-of-the-art methods in detecting facial landmarks.

Timing of the proposed approach directly depends on the number of branches and also the approach that we take to generate the manipulated faces. It is shown in [9] that semantically grouping the landmark locations increases the time of manipulated faces generation. However, it can overcome the problem of face distortion due to considering the semantic grouping. Therefore, there is a trade-off between the speed and accuracy of the proposed framework. However, in the case of aggregating with five branches and employing $GEAN_{Gadv}$ for generating manipulated faces, the framework runs in 17 FPS with NVIDIA TITAN X GPU.

4.2. Ablation Studies

Number of branches: In this section, we observe the effect of adding branches on the performance of the aggregated framework. We start with $k = 1$ in which there is no aggregation and one manipulated image is generated. We increase the number of branches from one to seven and measure the performance of aggregated network on the common, challenging, and full split of the 300-W dataset.

In addition, the number of branches in the training and testing phases is not necessarily the same. For example, the number of branches in the aggregated framework can be three while the number of branches in the testing phase is equal to 10. This is essentially important due to the time complexity of the framework during the training and testing

Train	Common test set				Challenging test set				Full test set				
	1	3	5	7	1	3	5	7	1*	1	3	5	7
1	4.40	3.77	3.48	3.43	5.44	5.35	5.30	5.28	4.80	4.80	4.49	4.07	4.02
3	3.67	3.25	3.03	2.98	5.33	5.27	5.18	5.10	4.68	4.46	4.01	3.77	3.74
5	3.35	2.99	2.68	2.66	5.26	4.97	4.71	4.67	4.63	4.04	3.64	3.05	3.01
7	3.32	2.93	2.65	2.63	5.22	4.90	4.65	4.60	4.59	3.96	3.56	3.00	2.97

Table 3: Comparison of NME on three test sets of 300-W with different numbers of branches for the training and testing. The column with asterisk demonstrates the results for evaluating the performance of our model without aggregation.

phases. In addition, one can train the network on two or three branches while test it on more branches to get more accurate results. Table 3 shows the evaluation results of 16 training and testing combinations, *i.e.*, four different training architectures ($K = 1, 3, 5, 7$) multiply four different testing architectures on 300-W common, challenging, and full test set, respectively.

As we can observe, the performance will be increased if the number of branches is increased during the training phase. However, we observe that adding more than five branches to the framework does not significantly improve the results with the cost of more computational complexity. The same behavior is observed for the testing framework. By increasing the number of branches in the testing phase, the accuracy is increased. This is useful when we want to reduce the computational complexity in training and maintaining the performance in the testing phase to some extent. Considering both accuracy and speed, we choose the framework with the number of training and testing branches equal to five for the sake of comparison with state-of-the-art (4.1).

We also conduct another experiment to demystify the effect of aggregation part in the proposed GEAN. In this case, GEAN with just one branch is trained on all the deformed and manipulated faces without the aggregation part. Table 3 and Figure 4 show the performance of $GEAN\ w/o\ Agg.$ compared to the proposed GEAN. For the sake of fair comparison, we trained the network on the same number of manipulated face images for both methods. By comparing column (1*) with column (1) of 300-W full test set, it is shown that the proposed GEAN which is trained with the exact same faces is superior to its counterpart without aggregation. Figure 4 also confirms the effectiveness of aggregation part and illustrates the fact that proposed GEAN performs beyond a careful augmentation.

A Comparison between Three Different Variations of GEAN: Three different approaches of $GEAN_{adv}$, $GEAN_{Gadv}$, and $GEAN_{GK}$ have been introduced in this paper. Through this section, we evaluate the performance of three different variations of our GEAN method on AFLW dataset. As Table 1 shows, both $GEAN_{Gadv}$ and $GEAN_{GK}$ outperform $GEAN_{adv}$ approach. We attribute this to the fact that $GEAN_{adv}$ does not consider grouping different landmarks semantically. This causes inconsistent displacements for the landmarks of one region (*e.g.*, left eye) and generate distorted images. In addition, the amount of displacement of landmarks in manipulated images might be greater than the manipulated images with the other two methods. However, this displacement might not be beneficial as it does not consider the general shape of each face region. Utilizing clipping constraint can mitigate this issue to some extent. However, this approach still suffers from not considering the semantic groups.

$GEAN_{Gadv}$ works the best among all three proposed approaches. Several reasons can explain this superiority. This approach considers the semantic relationship among the landmarks of same regions of the face. In addition, the manipulated images in this approach have different face IDs from the original face image. Therefore, in the framework with K branches, the aggregation is performed in K different face IDs. This makes this approach to preserve a reasonable relative distance among different groups of landmarks since it could fool the recognizer to misclassify it. However, this is not necessarily the case for the $GEAN_{GK}$ approach. This makes the $GEAN_{Gadv}$ to capture more important landmark displacement for the image manipulation which is beneficial for the aggregation. The advantages of the other two approaches (*i.e.*, adversarially attack technique in $GEAN_{adv}$ and semantic grouping of landmarks in $GEAN_{GK}$) is unified in $GEAN_{Gadv}$ which leads to a better landmark detection performance.

5. Conclusion

In this paper, we introduce a novel approach for facial landmark detection. The proposed method is an aggregated framework in which each branch of the framework contains a manipulated face. Three different approaches are employed to generate the manipulated faces and two of them perform the manipulation via the adversarial attacks to fool a face recognizer. This step can decouple from our framework and potentially used to enhance other landmark detectors [12, 31, 51]. Aggregation of the manipulated faces in different branches of GEAN leads to robust landmark detection. An ablation study is performed on the number of branches in training and testing phases and also on the effect different approaches of face image manipulation on the facial landmark detection. The results on the AFLW, 300-W, and COFW datasets show the superiority of our method compared to the state-of-the-art algorithms.

References

[1] and, C. C. Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4998–5006, June 2015.

[2] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013.

[3] F. L. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, June 1989.

[4] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d amp; 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1021–1030, Oct 2017.

[5] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1513–1520, 2013.

[6] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[7] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685, 2001.

[8] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[9] A. Dabouei, S. Soleymani, J. M. Dawson, and N. M. Nasrabadi. Fast geometrically-perturbed adversarial faces. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1979–1988, 2019.

[10] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1078–1085. IEEE, 2010.

[11] X. Dong, J. Huang, Y. Yang, and S. Yan. More is less: A more complicated network with less inference complexity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5840–5848, 2017.

[12] X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Style aggregated network for facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–388, 2018.

[13] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2385–2392, 2014.

[14] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. *arXiv preprint arXiv:1506.08347*, 2015.

[15] G. Ghiasi, C. C. Fowlkes, and C. Irvine. Using segmentation to predict the absence of occluded parts. In *BMVC*, pages 22–1, 2015.

[16] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[17] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[18] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015.

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, Oct 2017.

[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[21] S. Honari, J. Yosinski, P. Vincent, and C. Pal. Recombinator networks: Learning coarse-to-fine feature aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5743–5752, 2016.

[22] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[23] M. O. Irfanoglu, B. Gokberk, and L. Akarun. 3d shape-based face recognition using automatically registered facial surfaces. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 4, pages 183–186 Vol.4, Aug 2004.

[24] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu. Spatial transformer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2017–2025. Curran Associates, Inc., 2015.

[25] A. Jourabloo, M. Ye, X. Liu, and L. Ren. Pose-invariant face alignment with a single cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3219–3228, 2017.

[26] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, June 2014.

[27] 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 *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2144–2151, Nov 2011.

[28] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.

[29] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1):503–528, Aug 1989.

[30] F. Liu, D. Zeng, Q. Zhao, and X. Liu. Joint face alignment and 3d face reconstruction. In *European Conference on Computer Vision*, pages 545–560. Springer, 2016.

[31] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3691–3700, July 2017.

[32] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 483–499, Cham, 2016. Springer International Publishing.

[33] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas. A recurrent encoder-decoder network for sequential face alignment. In *European conference on computer vision*, pages 38–56. Springer, 2016.

[34] S. Reed, K. Sohn, Y. Zhang, and H. Lee. Learning to disentangle factors of variation with manifold interaction. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1431–II–1439. JMLR.org, 2014.

[35] S. E. Reed, Y. Zhang, Y. Zhang, and H. Lee. Deep visual analogy-making. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1252–1260. Curran Associates, Inc., 2015.

[36] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014.

[37] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment via regressing local binary features. *IEEE Transactions on Image Processing*, 25(3):1233–1245, 2016.

[38] 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.

[39] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[40] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[41] J. Su, Z. Wang, C. Liao, and H. Ling. Efficient and accurate face alignment by global regression and cascaded local refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[42] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[43] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[44] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016.

[45] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

[46] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4177–4187, 2016.

[47] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, June 2016.

[48] Y. Wu, C. Gou, and Q. Ji. Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3471–3480, 2017.

[49] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. X. Song. Spatially transformed adversarial examples. *CoRR*, abs/1801.02612, 2018.

[50] S. Xiao, J. Feng, L. Liu, X. Nie, W. Wang, S. Yan, and A. Kassim. Recurrent 3d-2d dual learning for large-pose facial landmark detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1642–1651, Oct 2017.

[51] J. Yang, Q. Liu, and K. Zhang. Stacked hourglass network for robust facial landmark localisation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2025–2033, July 2017.

[52] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[53] X. Yu, F. Zhou, and M. Chandraker. Deep deformation network for object landmark localization. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 52–70, Cham, 2016. Springer International Publishing.

[54] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European Conference on Computer Vision*, pages 1–16. Springer, 2014.

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

[56] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 94–108, Cham, 2014. Springer International Publishing.

[57] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):918–930, 2015.

[58] J. Zhao, M. Mathieu, R. Goroshin, and Y. Lecun. Stacked what-where auto-encoders. *arXiv preprint arXiv:1506.02351*, 2015.

[59] M. Zhu, D. Shi, M. Zheng, and M. Sadiq. Robust facial landmark detection via occlusion-adaptive deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3486–3496, 2019.

[60] S. Zhu, C. Li, C. C. Loy, and X. Tang. Unconstrained face alignment via cascaded compositional learning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3409–3417, June 2016.

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

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