Style Aggregated Network for Facial Landmark Detection

Xuanyi Dong[1], Yan Yan[1], Wanli Ouyang[2], Yi Yang[1]*
[1]University of Technology Sydney, [2] The University of Sydney
{xuanyi.dong,yan.yan-3}@student.uts.edu.au;
wanli.ouyang@sydney.edu.au; yi.yang@uts.edu.au

Abstract

Recent advances in facial landmark detection achieve success by learning discriminative features from rich deformation of face shapes and poses. Besides the variance of faces themselves, the intrinsic variance of image styles, e.g., grayscale vs. color images, light vs. dark, intense vs. dull, and so on, has constantly been overlooked. This issue becomes inevitable as increasing web images are collected from various sources for training neural networks. In this work, we propose a style-aggregated approach to deal with the large intrinsic variance of image styles for facial landmark detection. Our method transforms original face images to style-aggregated images by a generative adversarial module. The proposed scheme uses the style-aggregated image to maintain face images that are more robust to environmental changes. Then the original face images accompanying with style-aggregated ones play a duet to train a landmark detector which is complementary to each other. In this way, for each face, our method takes two images as input, i.e., one in its original style and the other in the aggregated style. In experiments, we observe that the large variance of image styles would degenerate the performance of facial landmark detectors. Moreover, we show the robustness of our method to the large variance of image styles by comparing to a variant of our approach, in which the generative adversarial module is removed, and no style-aggregated images are used. Our approach is demonstrated to perform well when compared with state-of-the-art algorithms on benchmark datasets AFLW and 300-W. Code is publicly available on GitHub: https://github.com/D-X-Y/SAN

1. Introduction

Facial landmark detection aims to detect the location of predefined facial landmarks, such as the corners of the eyes, eyebrows, the tip of the nose. It has drawn much attention recently as it is a prerequisite in many computer vision applications. For example, facial landmark detection can be applied to a large variety of tasks, including face recognition [74, 30], head pose estimation [58], facial reenactment [53] and 3D face reconstruction [28], to name a few.

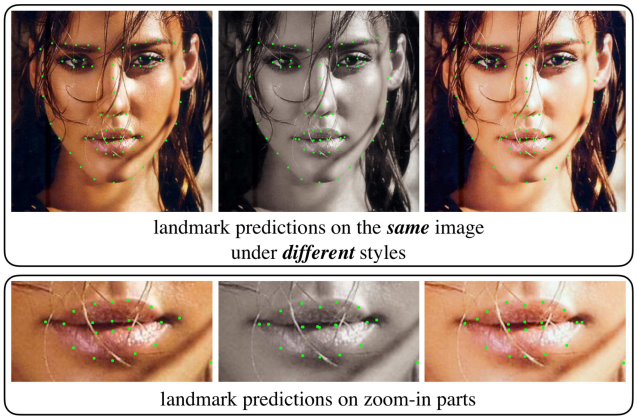


Figure 1. A face image in *three* different styles and the locations of the facial landmarks predicted by a facial landmark detector on them. The image styles, e.g., grayscale vs. color images, light vs. dark, intense vs. dull, can be quite distinct owing to various collection sources. The contents of the above three images are identical. The only difference is the image style. We apply a well-trained facial landmark detector to localize the facial landmarks. The zoom-in parts show the deviation among the predicted locations of the same facial landmarks on different styled images.

Recent advances in facial landmark detection mainly focus on learning discriminative features from abundant deformation of face shapes and poses, different expressions, partial occlusions, and others [58, 73, 59, 20]. A very typical framework is to construct features to depict the facial appearance and shape information by the convolutional neural networks (ConvNets) or hand-crafted features, and then learn a model, i.e., a regressor, to map the features to the landmark locations [64, 10, 7, 42, 72, 67, 40]. Most of them apply a cascade strategy to concatenate prediction modules and update the predicted locations of landmarks progressively [67, 10, 73].

*Corresponding author.

However, the issue from image style variation has been overlooked by recent studies on facial landmark detection. In real-world applications, face images collected in the wild usually are additionally under unconstrained variations [46, 73]. Large intrinsic variance of image styles, e.g., grayscale vs. color images, light vs. dark, intense vs. dull, is introduced when face images are collected under different environments and camera settings. The variation in image style causes the variation in prediction results. For example, Figure 1 shows *three* different styles of a face image and the facial landmark predictions on them when applying a well-trained detector. The contents of the three images are the same, but the visual styles are quite distinct, including original, grayscale and light. We can observe that the location predictions of a same facial landmark on them can be different. The zoom-in parts show the detailed deviation among the predicted locations of the same facial landmark on different styled images. This intrinsic variance of image styles would distort the prediction of the facial landmark detector and further degenerate the accuracy, which will be empirically demonstrated later. This problem commonly exists in the face in-the-wild landmark detection datasets [23, 46] (see Figure 2), and becomes inevitable for such face images captured under uncontrolled conditions.

Motivated by the issue of large variance of different image styles, we propose a Style-Aggregated Network (SAN) for facial landmark detection, which is insensitive to the large variance of image styles. The key idea of SAN is to first generate a pool of style-aggregated face images by the generative adversarial network (GAN) [16]. Then SAN exploits the complementary information from both the original images and the style-aggregated ones. The original images contain undistorted appearance contents of faces but may vary in image styles. The style-aggregated images contain stationary environments around faces, but may lack certain shape information due to the less fidelity caused by GAN. Therefore, our SAN takes both the original and style-aggregated faces together as complementary input, and applies a cascade strategy to generate the heatmap predictions which can be robust to the large variance of image styles.

To summarize, our contributions include:

1. To the best of our knowledge, we are the first to explicitly handle the problem caused by the variation of image styles in facial landmark detection problems, which has been overlooked in recent studies. We further empirically verify the performance degeneration caused by the large variance of image styles.

2. To facilitate style analysis, we release two new facial landmark detection datasets, 300W-Styles ($\approx$ 12000 images) and AFLW-Styles ($\approx$ 80000 images), by transferring the 300-W [46] and AFLW [23] into different styles.

Figure 2. Face samples from 300-W dataset. Different faces have different styles, whereas the style information may not be approachable in most facial landmark detection datasets.

3. We design a ConvNets architecture, i.e., Style-Aggregated Network (SAN), which exploits the mutual benefits of genuine appearance contents of faces and stationary environments around faces by simultaneously taking both original face images and style-unified ones.

4. In empirical studies, we verify the observation that the large variance of image styles would degenerate the performance of facial landmark detectors. Moreover, we show the insensitivity of SAN to the large variance of image styles and the state-of-the-art performance of SAN on benchmark datasets.

2. Related Work

2.1. Facial Landmark Detection

Increasing researchers focus on facial landmark detection [46]. The goal of facial landmark detection is to detect key-points in human faces, e.g., the tip of the nose, eyebrows, the eye corner and the mouth. Facial landmark detection is a prerequisite for a variety of computer vision applications. For example, Zhu et al. [74] take facial landmark detection results as input of 3D Morphable model. Wu et al. [58] propose a unified framework to deal with facial landmark detection, head pose estimation, and facial deformation analysis simultaneously, which couples each other. Thies et al. [53] use facial landmark detection confidences of keypoints in feature alignment for facial reenactment. Therefore, it is important to predict precise and accurate locations of the facial landmark.

A common approach to facial landmark detection problem is to learn a regression model [31, 64, 75, 5, 73, 7, 63]. Many of them leverage deep CNN to learn facial features and regressors in an end-to-end fashion [51, 31, 73] with a cascade architecture to progressively update the landmark estimation [73, 51, 10]. Yu et al. [66] propose a deep deformation network to incorporates geometric constraints within the CNN framework. Zhu et al. [73] leverage cascaded regressors to handle extreme head poses and rich shape deformation. Zhu et al. [72] utilize a coarse search over a shape space with diverse shapes to overcome the poor

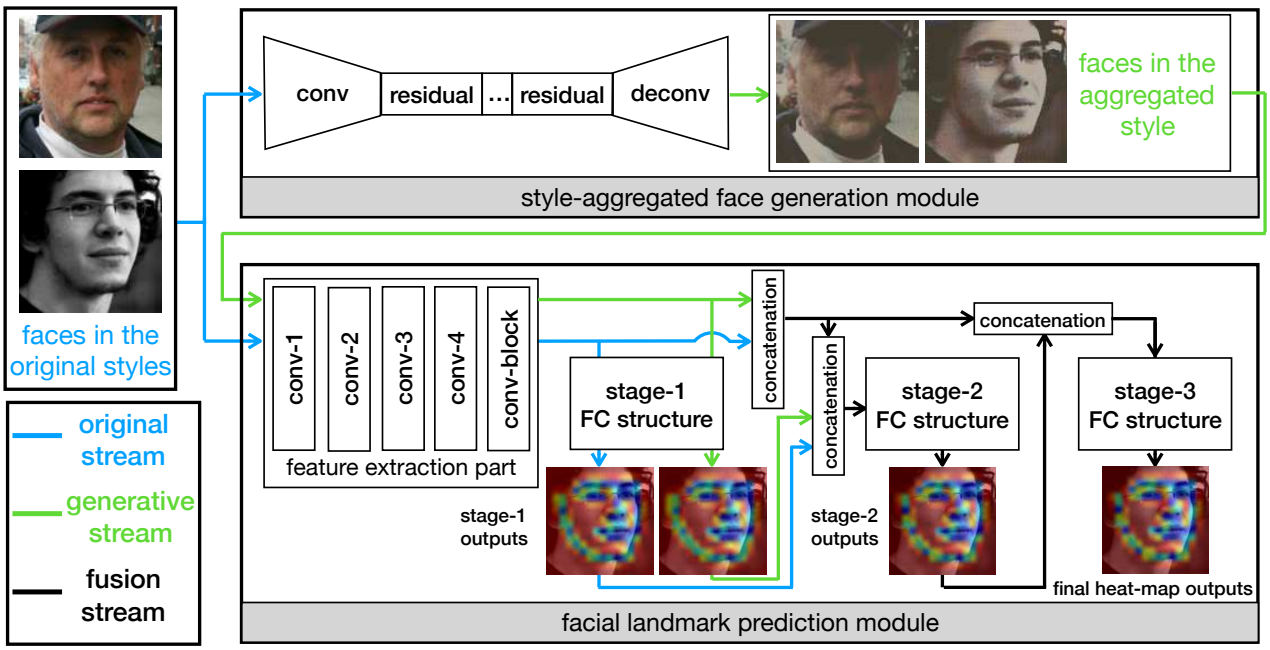


Figure 3. Overview of the SAN architecture. Our network consists of two components. The first is the style-aggregated face generation module, which transforms the input image into different styles and then combines them into a style-aggregated face. The second is the facial landmark prediction module. This module takes both the original image and the style-aggregated one as input to obtain two complementary features and then fuses the two features to generate heat-map predictions in a cascaded manner. "FC" means fully-convolution.

initialization problem. Lv et al. [31] present a deep regression architecture with two-stage reinitialization to explicitly deal with the initialization problem.

Another category of facial landmark detection methods takes the advantages of end-to-end training from deep CNN model to learn robust heatmap for facial landmark detection [27, 57, 6, 4]. Wei et al. [27] and Newell et al. [34] take the location with the highest response on the heatmap as the coordinate of the corresponding landmarks. Li et al. [27] enhance the facial landmark detection by multi-task learning. Bulat et al. [6] propose a robust network structure utilizing the state-of-the-art residual architectures.

These existing facial landmark detection algorithms usually focus on the facial shape information, e.g., the extreme head pose [20] or rich facial deformation [73]. However, few of them engage in a consideration of the intrinsic variance of image styles, e.g., grayscale vs. color images, light vs. dark and intense vs. dull. We also empirically demonstrate the performance fall caused by such intrinsic variance of image styles. This issue has been overlooked by recent studies but becomes inevitable as increasing web images are collected from various sources. Therefore, it is necessary to investigate the approach to dealing with the style variance, which is the focus of this paper.

Some researchers extend the landmark detection in the image to video settings [22, 13, 40] or 3D settings [6, 47]. In contrast, we focuses on image-based landmark detection.

2.2. Generative Adversarial Networks

We leverage the generator of trained GAN to generate faces into different styles to combat the large variance of face image styles.

GANs are first proposed in [16] to estimate generative models via an adversarial process. Following that, many researchers devoted great efforts to improve this research topic regarding theory [2, 8, 25, 35, 54] and applications [36, 41, 50, 71]. Some of them contribute to face applications, such as makeup-invariant face verification [26] and face aging [1]. In this work, we leverage a recently proposed technique, CycleGAN [71], to integrate a face generation model in our detection network. There are two different main focuses between this work and the previous works. First, we aim to group images into specific styles in an unsupervised manner, while they usually assume a stationary style in a dataset. Second, sophisticated face generation methods are not our target.

3. Methodology

How to design a neural network that is insensitive to the style variations for facial landmark detection? As illustrated in Figure 3, we design a network by combine two sub-modules to solve this problem: (1) The face generation module learns a neutral style of face images to combat the effect of style variations, i.e., transform faces with different styles into an aggregated style. (2) The landmark prediction

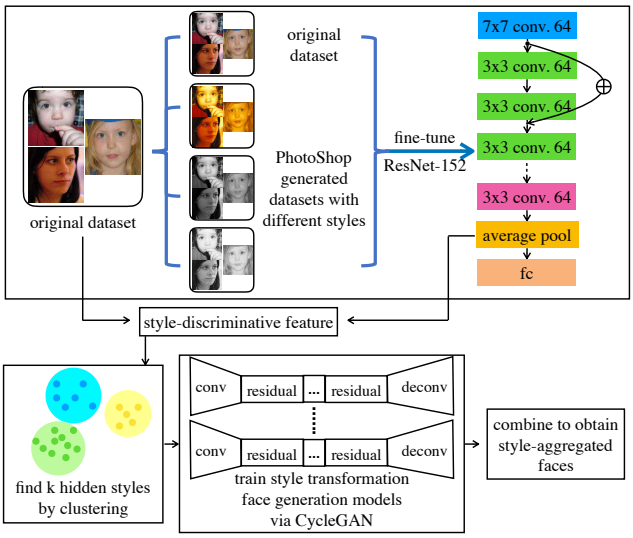


Figure 4. The pipeline to train the style-aggregated face generation module in an unsupervised way. We first utilize PS to transfer the original dataset into $C = 3$ different styles. These transferred datasets accompanying with the original dataset are then used to fine-tune the ResNet-152 with $C + 1$ classes. The fine-tuned features from the global average pooling layer can be considered as the style-discriminative features. We then leverage these features to cluster all images in the original dataset into k clusters, which can potentially contain the information of hidden styles. Lastly, we use these clustered data to train style transformation models via CycleGAN, and combine the trained models to obtain the final style-aggregated faces.

module leverages the complementary information from the neutral face and the original face to jointly predict the final coordinate for each landmark.

3.1. Style-Aggregated Face Generation Module

This module is motivated by the recent advances on image-to-image translation [19, 71] and style-transfer [14, 15, 56]. They can transform face images into a different style, whereas they require the style of images are already known in the training procedure as well as testing. However, face images in facial landmark detection datasets are usually collected from multiple sources. These images can have various styles, but we have no labels of these styles. Therefore, current facial landmark datasets do not align with the settings of image-to-image translation, and can thus not directly apply their techniques to our problem.

We design an unsupervised approach to learn a face generation model to first transfer faces into different styles and then combine them into an aggregated style. We first transfer the original dataset into three different styles by Adobe Photoshop (PS) [1]. These three transferred datasets accompanying with the original dataset are regarded as four classes to fine-tune the classification model [48, 17, 52, 11, 65, 62, 18]. The fine-tuned feature of the average-pooling layer thus has the style-discriminative characteristic, because the style information is learned in the training procedure by machine-generated style supervision.

To learn the stylized face generation model, we need to obtain the style information. For most face in-the-wild datasets, we can identify that faces have different styles. Figure 2 illustrates some examples of faces in various styles from 300-W [46]. However, it is hard to label such datasets with different styles due to two reasons: (1) Some style definitions are ambiguity, e.g., a face with light style can also be classified as the color. (2) It requires substantial labors to label the style information. Therefore, we leverage the learned style-discriminative feature to automatically cluster the whole dataset into k hidden styles by k-means.

Lastly, we regard the face images in different clusters as different hidden styles, and we then train face generation models to transfer styles via CycleGAN. CycleGAN is capable of preserving the structure of the input image because its cycle consistency loss guarantees the reconstructed images will match closely to the input images. The overall pipeline is illustrated in Figure 4. The final output is several face generation models that can transfer face images into different styles, and average the transferred faces into the style-aggregated ones.

3.2. Facial Landmark Prediction Module

The facial landmark prediction module leverages the mutual benefit of both the original images and the style-aggregated ones to overcome negative effects caused by style variations. This module is illustrated in Figure 3, where the green stream indicates the style-aggregated face and the blue stream represents the faces in the original styles. The blue stream contains undistorted appearance contents of faces but may vary in image styles. The green stream contains stationary environments around faces, but may lack certain shape information due to the less fidelity caused by GAN. By leveraging their complementary information, we can generate more robust predictions. The architecture is inspired by CPM [57]. We use the first four convolutional blocks from VGG-16 [49] followed by two additional convolution layers as feature extraction part. The feature extraction part takes the face image $\mathbf{I_o} \in \mathcal{R}^{h \times w}$ in the original styles and the one $\mathbf{I_s} \in \mathcal{R}^{h \times w}$ from the style-aggregated stream as input, where w and h represent the width and the height of image. In this part, each of the first three convolution blocks is followed by one pooling layer. It thus outputs the features $\mathbf{F} \in \mathcal{R}^{C \times h' \times w'}$ with eight times down-sample size compared to the input image $\mathbf{I}$, where $(h', w') = (h/8, w/8)$ and C is the channel of the last convolutional layer. The output features from the original and the style-aggregated faces are represented as $\mathbf{F_o}$ and $\mathbf{F_s}$, respectively. Three subsequent stages are used to produce 2D

[1] Three styles: Light, Gray and Sketch. See details in Sec 4.5.

belief maps [57]. Each stage is a fully-convolution structure. Its output tensor $\mathbf{H} \in \mathcal{R}^{(K+1)\times h' \times w'}$ has the same spatial size of the input tensor, where K indicates the number of landmarks. The first stage takes $\mathbf{F_o}$ and $\mathbf{F_s}$ as inputs and generate the belief maps for each of them, $\mathbf{H}_o$ and $\mathbf{H}_s$. The second stage g_2 takes the concatenation of $\mathbf{F_o}$, $\mathbf{F_s}$, $\mathbf{H}_o$ and $\mathbf{H}_s$ as inputs, and output the belief map for stage-2:

$$g_2(\mathbf{F}_o, \mathbf{F_s}, \mathbf{H}_o, \mathbf{H}_s) = \mathbf{H}_2. \quad (1)$$

The last stage is similar to the second one, which can be formulated as follows:

$$g_3(\mathbf{F}_o, \mathbf{F_s}, \mathbf{H}_2) = \mathbf{H}_3. \quad (2)$$

Following [34, 57], we minimize the following loss functions for each face image during the training procedure:

$$Loss = \sum_{i \in \{o,s,2,3\}} ||\mathbf{H}_i - \mathbf{H}_i^*||_F^2, \quad (3)$$

where $\mathbf{H}^*$ represents the ideal belief map.

To generate the final landmark coordinates, we first upsample the belief map $\mathbf{H}_3$ to the original image size using bicubic interpolation. We then use the argmax function on each belief map to obtain the coordinate of each landmark.

4. Experiments

4.1. Datasets

300-W [46]. This dataset annotates five face datasets with 68 landmarks, LFPW [3], AFW [75], HELEN [24], XM2VTS, IBUG. Following the common settings in [72, 31], we regard all the training samples from LFPW, HELEN and the full set of AFW as the training set, in which there is 3148 training images. 554 testing images from LFPW and HELEN form the common testing subset; 135 images from IBUG are regarded as the challenging testing subset. Both of these two subsets form the full testing set.

AFLW [23]. This dataset contains 21997 real-world images with 25993 faces in total. They provide at most 21 landmark coordinates for each face but excluding invisible landmark. Faces in AFLW usually have different pose, expression, occlusion or illumination, therefore causes difficulties to train a robust detector. Following the same setting as in [31, 73], we do not use the landmarks of two ears. There are two types of AFLW splits, AFLW-Full and AFLW-Frontal following [73]. AFLW-Full contains 20000 training samples and 4386 testing samples. AFLW-Front uses the same training samples as in AFLW-Full, but only use the 1165 samples with the frontal face as the testing set.

4.2. Experiment Settings

Training. We use PyTorch [39] for all experiments. To train the style-discriminative feature, we regard the original dataset and the PS-generated three datasets as four different classes. We then use them to fine-tune ResNet-152 ImageNet pre-trained model, and we train the model with the learning rate of 0.01 for two epochs in total. We use k-means to cluster the whole dataset into $k = 3$ groups, and regard the group with the maximum element and the group with the minimum as two different style sets by default. These two different groups are then used to train our style-unified face generation module via Cycle-GAN [71]. We follow the similar training settings as in [71], whereas we train our model with the batch size of 32 on two GPUs, and also set the identity loss in [71] as 0.1. To train the facial landmark prediction module, the first four convolutional blocks are initialized by VGG-16 ImageNet pre-trained model, and other layers are initialized using a Gaussian distribution with the variance of 0.01. Lastly, we train the facial landmark prediction model with the batch size of 8 and weight decay of 0.0005 on two GPUs. We start the learning rate at 0.00005 and reduce the learning rate at 30th/35th/40th/45th epochs by 0.5, and we then stop training at 50th epoch. The face bounding box is expanded by the ratio of 0.2. We use the random crop for pre-processing during training as data argumentation.

Method	Common	Challenging	Full Set
SDM [64]	5.57	15.40	7.52
ESR [7]	5.28	17.00	7.58
LBF [43]	4.95	11.98	6.32
CFSS [72]	4.73	9.98	5.76
MDM [55]	4.83	10.14	5.88
TCDCN [68]	4.80	8.60	5.54
Two-Stage$_{OD}$ [31]	4.36	7.56	4.99
Two-Stage$_{GT}$ [31]	4.36	7.42	4.96
RDR [61]	5.03	8.95	5.80
Pose-Invariant[20]	5.43	9.88	6.30
SAN$_{OD}$	3.41	7.55	4.24
SAN$_{GT}$	**3.34**	**6.60**	**3.98**

Table 1. Normalized mean errors (NME) on 300-W dataset.

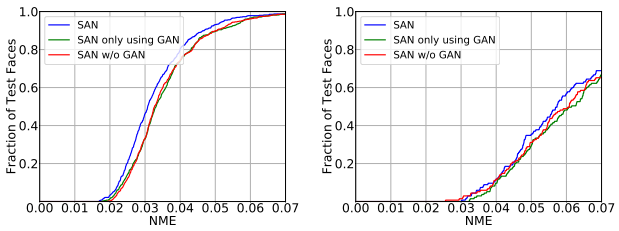
Evaluation. Normalized Mean Error (NME) is usually applied to evaluate the performance for facial landmark predictions [31, 43, 73]. For 300-W dataset, we use the interocular distance to normalize mean error following the same setting as in [46, 31, 7, 43]. For AFLW dataset, we use the face size to normalize mean error [31]. We also use Cumulative Error Distribution (CED) curve to compare the algorithms provided in [45]. Area Under the Curve (AUC) @ 0.08 error is also employed for evaluation [6, 55].

4.3. Comparison with State-of-the-art Methods

Results on 300-W. Table 1 shows the performance of different facial landmark detection algorithms on the 300-W. We compare our approach with recently proposed state-of-the-art algorithms [31, 61, 20]. We compare our approaches based on two types of face bounding boxes: (1) ground truth bounding box, denoted as GT; (2) official detector, denoted as OD. SAN achieves very competitive re-

Methods	SDM [64]	ERT [21]	LBF [43]	CFSS [72]	CCL [73]	Two-Stage [31]	SAN
AFLW-Full	4.05	4.35	4.25	3.92	2.72	2.17	1.91
AFLW-Front	2.94	2.75	2.74	2.68	2.17	-	1.85

Table 2. Comparisons of normalized mean (NME) errors on AFLW dataset.



(a) 300-W Common Testing Set (b) 300-W Challenging Testing Set

Figure 5. CED curves for 300-W common and challenging testing sets. The blue line shows the performance of SAN. The green and red lines indicate SAN with only the style-aggregated face and with only the original face being the input, respectively.

Figure 6. Qualitative results of the clustered face images from 300-W by using the style-discriminative features. The face images in each cluster have some different hidden styles. For example, the first cluster has many grayscale faces; the second cluster shows the dark illumination; the last cluster shows the light illumination. We generate the mean face for each cluster. These mean face images show the very *similar* face, while they have quite *different* environments.

sults compared with others by using the same face bounding box (OD). We improve the performance of NME on 300-W common set by relative 21.8% compared to the state-of-the-art method. It can further enhance our approach by applying a better initialization (GT). This implies that SAN has potential to be more robust by incorporating the face alignment [31] or landmark refinement [73, 55] methods.

Results on AFLW. We use the training/testing splits and the bounding box provided from [73, 72]. Table 2 shows the performance comparison on AFLW. Our SAN also achieves the very competitive NME results, which are better than the previous state-of-the-art by more than 11% on AFLW-Full. On the AFLW-Front testing set, our result is also better than state-of-the-art by more than 14%. We find that more clusters and more generation models in style-aggregated face generation module will obtain a similar result as $k = 3$, we thus use the setting of $k = 3$ by default.

SAN achieves new state-of-the-art results on two benchmark datasets, e.g., 300-W and AFLW. It takes two complementary images to generate predictions which are insensitive to style variations. The idea of using the two-stream input for facial landmark detection can be complementary to other algorithms [20, 31, 61, 73]. They usually do not consider the effect of image style, while the style-aggregated face in the two-steam input can handle this problem.

4.4. Ablation Studies

In this section, we first verify the significance of each component in our proposed SAN. Figure 5 shows the comparison regarding CED curves for our SAN and two variants of SAN on the 300-W common and testing sets. As we can observe, the performance will significantly be deteriorated if we remove the original face image or the generated style-aggregated face image. This observation demonstrates that taking two complementary face images as the input benefits the facial landmark prediction results.

Figure 6 shows the results of k-means clustering on 300-W dataset. 300-W dataset is the face in-the-wild dataset, where face images have large style variations but this style information is not approachable. Our style-discriminative feature is capable of distinguishing images with different hidden styles. We can find that most of the face images in one cluster share a similar style. The mean face images generated from three clusters contain different styles. If we directly use ImageNet pre-trained features for k-means clustering, we can not guarantee to group faces into different hidden styles. In experiments, we find that ImageNet pre-trained features tend to group face images by the gender or other information.

4.5. Discussions of Benchmark Datasets

Facial landmark detection datasets with constrained face images [33] usually have the similar environment for each image. There are only small style changes in these datasets, and they may also not be applicable for real-world applications due to the small face variance. We thus do not discuss these datasets in this paper. The face in-the-wild datasets [46, 23] contain face images with large intrinsic variance. However, this intrinsic variance information is not available from the official datasets, but can also affect the predictions of the detector. Therefore, we propose two

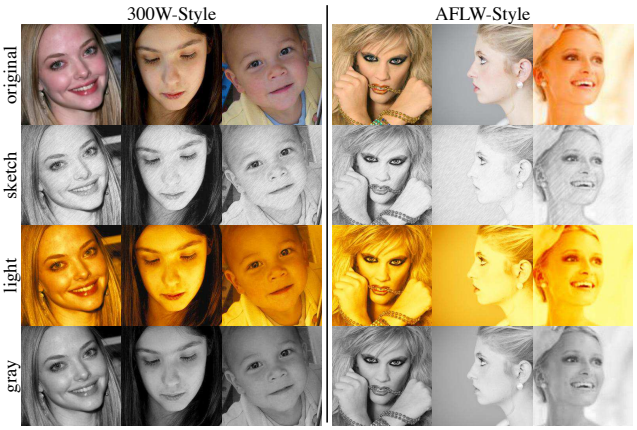


Figure 7. Our PS-generated datasets based on 300-W and AFLW with the original and three synthetic styles, i.e., sketch, light and gray. These datasets have different styles and can be used to facilitate style analysis.

new datasets, 300W-Style and AFLW-Style, to facilitate the style analysis for facial landmark detection problem.

As shown in Figure 7, 300W-Style consists of four different styles, original, sketch, light and gray. The original part is the original 300-W datasets, and the other three synthetic styles are generated using PS. Each image in 300W-Style is corresponding to one image in the 300-W dataset, and we thus directly use the annotation provided from 300-W for our 300W-Style. AFLW-Style is similar as 300W-Style, which transfer the AFLW dataset into three different styles. For training and testing split, we follow the common settings of the original datasets [46, 23].

Can PS-generated images be realistic? Internet users usually use PS (or similar software) to change image styles and/or edit image content; thus PS-generated images are indeed realistic in many real-world applications. In addition, we have chosen three representative filters to generate images of different styles. These filters have been widely used by users to edit their photos and upload to the Internet. Therefore, the proposed datasets are realistic.

Test \ Train	Original	Light	Gray	Sketch
SAN w/o GAN				
Original	3.37	3.56	3.77	3.92
Light	3.61	3.41	4.01	4.13
Gray	3.47	3.79	3.43	3.60
Sketch	3.71	3.97	3.66	3.40
SAN				
Original	3.34 (↑ 0.8%)	3.44 (↑ 3.3%)	3.46 (↑ 8.2%)	3.54 (↑ 9.7%)
Light	3.48 (↑ 3.6%)	3.39 (↑ 0.5%)	3.56 (↑ 11.2%)	3.68 (↑ 10.9%)
Gray	3.45 (↑ 0.6%)	3.56 (↑ 6.1%)	3.38 (↑ 1.4%)	3.52 (↑ 2.2%)
Sketch	3.53 (↑ 4.9%)	3.62 (↑ 8.8%)	3.55 (↑ 3.0%)	3.35 (↑ 1.4%)

Table 3. Comparisons of NME on the 300W-Style common testing set. We use different styles for training and testing.

Test \ Train	Original	Light	Gray	Sketch
SAN w/o GAN				
Original	6.88	7.82	7.84	7.74
Light	7.31	7.16	8.91	8.67
Gray	7.08	8.59	6.77	6.98
Sketch	7.59	8.68	7.17	6.83
SAN				
Original	6.60 (↑ 4.1%)	7.00 (↑ 10.5%)	6.73 (↑ 14.2%)	6.97 (↑ 9.9%)
Light	7.15 (↑ 2.2%)	7.08 (↑ 1.1%)	7.26 (↑ 18.5%)	7.15 (↑ 17.5%)
Gray	6.91 (↑ 2.4%)	7.18 (↑ 16.4%)	6.69 (↑ 1.1%)	6.97 (↑ 0.2%)
Sketch	7.08 (↑ 6.7%)	7.64 (↑ 12.0%)	6.95 (↑ 3.1%)	6.77 (↑ 0.8%)

Table 4. Comparisons of NME on the 300W-Style challenging testing set. We use different styles for training and testing.

Test \ Train	Original	Light	Gray	Sketch
SAN w/o GAN				
Original	4.06	4.39	4.57	4.67
Light	4.33	4.14	4.97	5.02
Gray	4.19	4.73	4.08	4.26
Sketch	4.47	4.89	4.35	4.07
SAN				
Original	3.98 (↑ 1.9%)	4.14 (↑ 5.7%)	4.10 (↑ 10.2%)	4.21 (↑ 9.9%)
Light	4.20 (↑ 3.0%)	4.12 (↑ 0.4%)	4.29 (↑ 13.7%)	4.36 (↑ 13.1%)
Gray	4.13 (↑ 1.4%)	4.27 (↑ 9.7%)	4.03 (↑ 1.2%)	4.20 (↑ 1.4%)
Sketch	4.23 (↑ 5.4%)	4.41 (↑ 6.7%)	4.21 (↑ 3.2%)	4.02 (↑ 1.2%)

Table 5. Comparisons of NME on the 300W-Style full testing set. We use different styles for training and testing.

Effect of SAN for style variances. These two proposed datasets can be used to analyze the effect of face image styles for facial landmark detection. We consider the situation that testing set has a different style with the training set. For example, we train the detector on the light-style 300-W training set and evaluate the well-trained detector on 300-W testing sets with different styles. Table 3, Table 4 and Table 5 show the evaluation results of 16 training and testing style combinations, i.e., four different training styles multiply four different testing styles. Our SAN algorithm is specifically designed to deal with style variances for face landmark detection. When style variance between the

Figure 8. Representative results on 300-W. The red points in the first line indicate the ground-truth landmarks. The blue points in the second line and the green points in the third line indicate the landmark predictions from the base detector and SAN, respectively.

training and testing sets is large (e.g., light and gray), our approach usually obtains a significant improvement. However, if style variance between the training and testing sets is not that large (e.g., gray and sketch), the improvement of SAN is less significant. On average, SAN obtains 7% relative improvement on the full testing set of the 300W-Style dataset when the training style is different from the testing style. Moreover, our SAN achieves consistent improvements over all the 16 different train-test style combinations. This demonstrates the effectiveness of our method.

Self-Evaluation: We compare two variants of our SAN: (1) train SAN without GAN using the training set of AFLW-Style and the testing set of AFLW. This can be considered as data argumentation, because the amount of training data that we use is four times larger than the original one. In this case, our SAN can achieve 79.82 AUC@0.08 on AFLW-Full by only using the original AFLW training set, while the data argumentation one achieves a worse performance, 78.99 AUC@0.08, than SAN. SAN is better than the data argumentation way, which uses our PS-generated images as additional training data. (2) replace the style-aggregated stream of SAN by a Photo-generated face image. If we train the detector on the original style 300-W training set and test it on the gray style 300-W challenging test set, our SAN can achieve 6.91 NME. However, replacing the style-aggregated stream by light style images can only achieve 7.30 NME, which is worse than ours. SAN can always achieve better results than the replaced variant, except for replacing the style-aggregated stream by the testing style. SAN can automatically learn the hidden styles in the dataset and generate the style-aggregated face images. This automatic way is better than providing images with a fixed style.

Error Analysis: The faces in uncontrolled conditions have large variations regarding the image style. Detectors will usually fail when image style changes a lot, whereas our SAN is insensitive to this style change. Figure 8 shows the qualitative results of our SAN and the base detector on 300-W. The first line shows the ground truth landmarks. The second and third lines show the predictions from SAN without GAN and SAN, respectively. In the first column, the base detector fails for the predictions on the face contour, while the predictions from SAN still preserves the overall structure. In the fourth column, some perdition from the base detector drifts to the right, while SAN not.

5. Conclusion & Future Work

The large intrinsic variance of image styles, which comes from their uncontrolled collection sources, has been overlooked by recent studies in facial landmark detection. To deal with this issue, we propose a style-aggregated network (SAN). SAN takes two complementary images for each face, one in the original style and the other in the aggregated style that is generated by GAN. Empirical studies verify that style variations degenerate the performance of landmark detection, and SAN is robust to the large variance of image styles. Additionally, SAN achieves state-of-the-art performance on 300-W and AFLW datasets.

The first step of SAN is to generate the style-aggregated images. This step can be decoupled from our landmark detector, and potentially used to improve other landmark detectors [7, 43, 72, 68, 37]. Moreover, the intrinsic variance of image styles also exists in other computer vision tasks, such as object detection [12, 44, 38, 9, 29] and person re-identification [60, 69, 70, 32]. Therefore, the style-aggregation method can also be used to solve the problem of the style variance in other applications. In our future work, we will explore how to generalize the style-aggregation method for other computer vision tasks.

Acknowledgment. Yi Yang is the recipient of a Google Faculty Research Award. Wanli Ouyang is supported by SenseTime Group Limited. We acknowledge the Data to Decisions CRC (D2D CRC) and the Cooperative Research Centres Programme for funding this research.

References

[1] G. Antipov, M. Baccouche, and J.-L. Dugelay. Face aging with conditional generative adversarial networks. In *ICIP*, 2017. 3

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 3

[3] 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. 5

[4] A. Bulat and G. Tzimiropoulos. Convolutional aggregation of local evidence for large pose face alignment. In *BMVC*, 2016. 3

[5] A. Bulat and G. Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *ICCV*, 2017. 2

[6] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017. 3, 5

[7] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014. 1, 2, 5, 8

[8] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016. 3

[9] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: Object detection via region-based fully convolutional networks. In *NIPS*, 2016. 8

[10] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *CVPR*, 2010. 1, 2

[11] X. Dong, J. Huang, Y. Yang, and S. Yan. More is less: A more complicated network with less inference complexity. In *CVPR*, 2017. 4

[12] X. Dong, D. Meng, F. Ma, and Y. Yang. A dual-network progressive approach to weakly supervised object detection. In *ACM Multimedia*, 2017. 8

[13] X. Dong, S.-I. Yu, X. Weng, S.-E. Wei, Y. Yang, and Y. Sheikh. Supervision-by-Registration: An unsupervised approach to improve the precision of facial landmark detectors. In *CVPR*, 2018. 3

[14] V. Dumoulin, J. Shlens, M. Kudlur, A. Behboodi, F. Lemic, A. Wolisz, M. Molinaro, C. Hirche, M. Hayashi, E. Bagan, et al. A learned representation for artistic style. In *ICLR*, 2017. 4

[15] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 4

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 2, 3

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[18] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. In *CVPR*, 2017. 4

[19] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 4

[20] A. Jourabloo, X. Liu, M. Ye, and L. Ren. Pose-invariant face alignment with a single cnn. In *ICCV*, 2017. 1, 3, 5, 6

[21] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014. 6

[22] M. H. Khan, J. McDonagh, and G. Tzimiropoulos. Synergy between face alignment and tracking via discriminative global consensus optimization. In *ICCV*, 2017. 3

[23] M. Koestinger, 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 *ICCV-W*, 2011. 2, 5, 6, 7

[24] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*, 2012. 5

[25] Y. Li, A. Schwing, K.-C. Wang, and R. Zemel. Dualing GANs. In *NIPS*, 2017. 3

[26] Y. Li, L. Song, X. Wu, R. He, and T. Tan. Anti-Makeup: Learning a bi-level adversarial network for makeup-invariant face verification. In *AAAI*, 2018. 3

[27] Y. Li, B. Sun, T. Wu, and Y. Wang. Face detection with end-to-end integration of a convnet and a 3d model. In *ECCV*, 2016. 3

[28] F. Liu, D. Zeng, Q. Zhao, and X. Liu. Joint face alignment and 3D face reconstruction. In *ECCV*, 2016. 1

[29] Y. Liu, H. Li, J. Yan, F. Wei, X. Wang, and X. Tang. Recurrent scale approximation for object detection in cnn. In *ICCV*, 2017. 8

[30] Y. Liu, F. Wei, J. Shao, L. Sheng, J. Yan, and X. Wang. Exploring disentangled feature representation beyond face identification. In *CVPR*, 2018. 1

[31] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou. A deep regression architecture with two-stage reinitialization for high performance facial landmark detection. In *CVPR*, 2017. 2, 3, 5, 6

[32] F. Ma, D. Meng, Q. Xie, Z. Li, and X. Dong. Self-paced co-training. In *ICML*, 2017. 8

[33] S. Milborrow, J. Morkel, and F. Nicolls. The MUCT landmarked face database. *Pattern Recognition Association of South Africa*, 201(0), 2010. 6

[34] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 3, 5

[35] S. Nowozin, B. Cseke, and R. Tomioka. F-GAN: Training generative neural samplers using variational divergence minimization. In *NIPS*, 2016. 3

[36] A. Osokin, A. Chessel, R. E. C. Salas, and F. Vaggi. GANs for biological image synthesis. In *ICCV*, 2017. 3

[37] W. Ouyang, X. Chu, and X. Wang. Multi-source deep learning for human pose estimation. In *CVPR*, 2014. 8

[38] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *CVPR*, 2012. 8

[39] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. In *NIPS-W*, 2017. 5

[40] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas. A recurrent encoder-decoder network for sequential face alignment. In *ECCV*, 2016. 1, 3

[41] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 3

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

[43] 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. 5, 6, 8

[44] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 8

[45] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18, 2016. 5

[46] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV-W*, 2013. 2, 4, 5, 6, 7

[47] T. Simon, J. Valmadre, I. Matthews, and Y. Sheikh. Kronecker-Markov prior for dynamic 3D reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2201–2214, 2017. 3

[48] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4

[49] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4

[50] L. Sixt, B. Wild, and T. Landgraf. RenderGAN: Generating realistic labeled data. In *ICLR*, 2017. 3

[51] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013. 2

[52] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 4

[53] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016. 1, 2

[54] I. Tolstikhin, S. Gelly, O. Bousquet, C.-J. Simon-Gabriel, and B. Schölkopf. AdaGan: Boosting generative models. In *NIPS*, 2017. 3

[55] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *CVPR*, 2016. 5, 6

[56] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, 2016. 4

[57] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 3, 4, 5

[58] Y. Wu, C. Gou, and Q. Ji. Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. In *CVPR*, 2017. 1, 2

[59] Y. Wu and Q. Ji. Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In *CVPR*, 2016. 1

[60] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*, 2018. 8

[61] 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 *CVPR*, 2017. 5, 6

[62] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 4

[63] J. Xing, Z. Niu, J. Huang, W. Hu, and S. Yan. Towards multi-view and partially-occluded face alignment. In *CVPR*, 2014. 2

[64] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013. 1, 2, 5, 6

[65] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, and D. Xu. Image classification by cross-media active learning with privileged information. *IEEE Transactions on Multimedia*, 18(12):2494–2502, 2016. 4

[66] X. Yu, F. Zhou, and M. Chandraker. Deep deformation network for object landmark localization. In *ECCV*, 2016. 2

[67] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *ECCV*, 2014. 1

[68] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014. 5, 8

[69] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 8

[70] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018. 8

[71] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 3, 4, 5

[72] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, 2015. 1, 2, 5, 6, 8

[73] S. Zhu, C. Li, C.-C. Loy, and X. Tang. Unconstrained face alignment via cascaded compositional learning. In *CVPR*, 2016. 1, 2, 3, 5, 6

[74] 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 *CVPR*, 2015. 1, 2

[75] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012. 2, 5