

Aggregation via Separation: Boosting Facial Landmark Detector with Semi-Supervised Style Translation

Shengju Qian¹, Keqiang Sun², Wayne Wu^{2,3}, Chen Qian³, Jiaya Jia^{1,4}

¹The Chinese University of Hong Kong ²Tsinghua University

³SenseTime Research ⁴YouTu Lab, Tencent

{sjqian, leo.jia}@cse.cuhk.edu.hk, skq17@mails.tsinghua.edu.cn, {wuwenyan, qianchen}@sensetime.com

Abstract

Facial landmark detection, or face alignment, is a fundamental task that has been extensively studied. In this paper, we investigate a new perspective of facial landmark detection and demonstrate it leads to further notable improvement. Given that any face images can be factored into space of style that captures lighting, texture and image environment, and a style-invariant structure space, our key idea is to leverage disentangled style and shape space of each individual to augment existing structures via style translation. With these augmented synthetic samples, our semi-supervised model surprisingly outperforms the fully-supervised one by a large margin. Extensive experiments verify the effectiveness of our idea with state-of-the-art results on WFLW [69], 300W [56], COFW [7], and AFLW [36] datasets. Our proposed structure is general and could be assembled into any face alignment frameworks. The code is made publicly available at <https://github.com/thesouthfrog/stylealign>.

1. Introduction

Facial landmark detection is a fundamentally important step in many face applications, such as face recognition [44], 3D face reconstruction [17], face tracking [33] and face editing [61]. Accurate facial landmark localization was intensively studied with impressive progress made in these years. The main streams are learning a robust and discriminative model through effective network structure [69], usage of geometric information [6, 31], and correction of loss functions [19].

It is common wisdom now that factors such as variation of expression, pose, shape, and occlusion could greatly affect performance of landmark localization. Almost all prior work aims to alleviate these problems from the perspective of structural characteristics, such as disentangling 3D pose to provide shape constraint [37], and utilizing dense bound-



Figure 1: Problem in a well-trained facial landmark detector. It is biased towards unconstrained environment factors, including lighting, image quality, and occlusion. We regard these degradations as “style” in our analysis.

ary information [69]. The influence of “environment” still lacks principled discussion beyond structure. Also, considering limited labeled data for this task, how to optimally utilize limited training samples remains unexplored.

About “environment” effect, distortion brought by explicit image style variance was observed recently [15]. We instead utilize style transfer [30, 21] and disentangled representation learning [62, 10, 42, 16, 25] to tackle the face alignment problem, since style transfer aims at altering style while preserving content. In practice, image content refers to objects, semantics and sharp edge maps, whereas style could be color and texture.

Our idea is based on the purpose of facial landmark detection, which is to regress “facial content” – the principal component of facial geometry – by filtering unconstrained “styles”. The fundamental difference to define “style” from that of [15] is that we refer it to image background, lighting, quality, existence of glasses, and other factors that prevent detectors from recognizing facial geometry. We note every face image can be decomposed into its facial structure

along with a distinctive attribute. It is a natural conjecture that *face alignment could be more robust if we augment images only regarding their styles*.

To this end, we propose a new framework to augment training for facial landmark detection without using extra knowledge. Instead of directly generating images, we first map face images into the space of structure and style. To guarantee the disentanglement of these two spaces, we design a conditional variational auto-encoder [35] model, in which Kullback-Leiber (KL) divergence loss and skip connections are incorporated for compact representation of style and structure respectively. By factoring these features, we perform visual style translation between existing facial geometry. Given existing facial structure, faces with glasses, of poor quality, under blur or strong lighting are *re-rendered* from corresponding style, which are used to further train the facial landmark detectors for a rather general and robust system to recognize facial geometry.

Our main contribution is as follows.

1. We offer a new perspective for facial landmark localization by factoring style and structure. Consequently, a face image is decomposed and rendered from distinctive image style and facial geometry.
2. A novel semi-supervised framework based on conditional variational auto-encoder is built upon this new perspective. By disentangling style and structure, our model generates style-augmented images via style translation, further boosting facial landmark detection.
3. We propose a new dataset based on AFLW [36] with new 68-point annotation. It provides challenging benchmark considering large pose variation.

With extensive experiments on popular benchmark datasets including WFLW [69], 300W [56], COFW [7] and AFLW [36], our approach outperforms previous state-of-the-arts by a large margin. It is general to be incorporated into various frameworks for further performance improvement. Our method also works well under limited training computation resource.

2. Related Work

This work has close connection with the areas of facial landmark detection, disentangled representation and self-supervised learning.

Facial Landmark Detection This area has been extensively studied over past years. Classic parameterized methods, such as active appearance models (AAMs) [11, 57, 47, 32] and constrained local models (CLMs) [12] provide satisfying results. SDM [73], cascaded regression, and their variants [67, 83, 82, 8, 7, 9, 73, 64, 18] were also proposed.

Recently, with the power of deep neural networks, regression-based models are able to produce better results. They are mainly divided into two streams of direct coordinate regression [80, 45, 63, 50] and heatmap-based regression [51, 5, 13, 75, 48]. Meanwhile, in [80], auxiliary attributes were used to learn a discriminative representation. Recurrent modules [63, 72, 52] were introduced then. Lately, methods improved performance via semi-supervised learning [26]. Influence of style variance was also discussed in [15], where a style aggregated component provides a stationary environment for landmark detector. Our solutions are distinct with definition of “style”, different from prior work. Our solution does not rely on the aggregation architecture, and instead is based on a semi-supervised scheme.

Disentangled Representation Our work is also related to disentangled representation learning. Disentanglement is necessary to control and further alter the latent information in generated images. Under the unsupervised setting, InfoGAN [10] and MINE [3] learned disentangled representation by maximizing the mutual information between latent code and data observation. Recently, image-to-image translation [43, 28, 42, 29] explored the disentanglement between style and content without supervision. In structured tasks such as conditional image synthesis [46], keypoints [16, 53] and person mask [2] were utilized as self-supervision signals to disentangle factors, such as foreground, background and pose information. As our “style” is more complex while “content” is represented by facial geometry, traditional style transfer [21] is inapplicable since it may suffer from structural distortion. In our setting, by leveraging the structure information base on landmarks, our separation component extracts the style factor from each face image.

Self-Supervised Learning Our method also connects to self-supervised learning. The mainstream work, such as [76], directly uses image data to provide proxy supervision through multi-task feature learning. Another widely-adopted approach is to use video data [66]. Visual invariance of the same instance could be captured in a consecutive sequence of video frames [20, 68, 40, 85, 60, 59, 66]. Also, there is work focusing on fixed characteristics of objects from data statistics [14, 78, 79, 38, 39], such as image patch level information [14]. These methods learn visual invariance, which could essentially provide a generalized feature of objects.

Our landmark localization involves computing the visual invariance. But our approach is different from prior self-supervised frameworks. Our goal lies in extracting facial structure and keypoints considering different environment factors, including occlusion, lighting, makeup and so on. Eliminating the influence of style makes it possible to reliably alter or process face structure and accordingly recog-

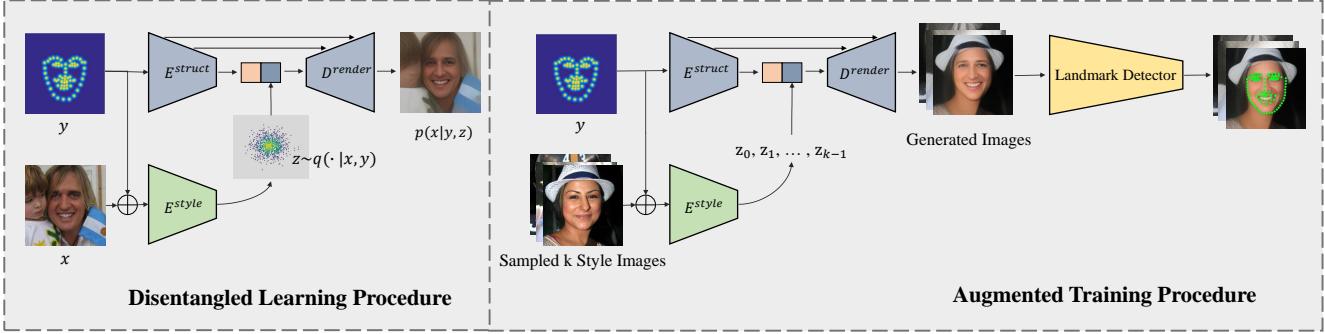


Figure 2: Our framework. It consists of two stages. The first stage is to train the network to disentangle face images to style and structure space. At the second stage, style translation is performed to augment training of facial landmark detectors.

nize invariant features. It thus better deals with style variation, which commonly exists in natural images.

3. Proposed Framework

Our framework consists of two parts. One learns the disentangled representation of facial appearance and structure, while the other can be any facial landmark detectors. As illustrated in Fig. 2, during the first phase, conditional variational auto-encoder is proposed for learning disentangled representation between style and structure. In the second phase, after translating style from other faces, “stylized” images with their structures are available for boosting training performance and our style-invariant detectors.

3.1. Learning Disentangled Style and Structure

Given an image x , and its corresponding structure y . Two essential descriptors of a face image are facial geometry and image style. Facial geometry is represented by labeled landmarks, while style captures all environmental factors that are mostly implicit, as described above. With this setting, if the latent space of style and shape is mostly uncorrelated, using Cartesian product of z and y latent space should capture all variation included in a face image. Therefore, the generator that re-renders a face image based on style and structure can be modeled as $p(x|y, z)$.

To encode the style and structure information and compute the parametric distribution $p(x|y, z)$, a conditional variational auto-encoder based network, which introduces two encoders, is applied. Our network consists of a structure estimator E^{struct} to encode landmark heatmaps into structure latent space, a style encoder E^{style} that learns the style embedding of images, and a decoder that re-renders the style and structure to image space.

As landmarks available in this task, the facial geometry is represented by stacking landmarks to heat maps. Our goal therefore becomes inferring disentangled style code z from a face image and its structure by maximizing the conditional

likelihood of

$$\log p(x|y) = \log \int_z p(x, z|y) dz \geq \mathbb{E}_q[\log \frac{p(x, z|y)}{q(z|x, y)}] \quad (1)$$

$$= \mathbb{E}_q[\log p(x|z, y)] - D_{KL}[q(z|x, y), p(z|y)].$$

In particular, the generator G_θ^{full} contains two encoders and a decoder (renderer), i.e., E_ϕ^{style} , E^{struct} and D_{render} , where G_θ^{full} and E_ϕ^{style} respectively estimate parameters of $p(x|y, z)$ and $q(z|x, y)$. Consequently, the full loss function on learning separating information of style and structure is written as

$$\mathcal{L}_{disentangle}(x, \theta, \phi) = -KL(q_\phi(z|x, y))||p_\theta(z|y)) \quad (2)$$

$$+ \mathcal{L}_{rec}(x, G^{full}(E^{style}(x, y), E^{struct}(y))).$$

KL-Divergence Loss Kullback-Leiber (KL) divergence loss severs as a key component in our design to help the encoder to learn decent representation. Basically, the KL-divergence measures the similarity between the variational posterior and prior distribution. In our framework, it is taken as regularization that discourages E^{style} to encode structure-related information. As the prior distribution is commonly assumed to be a unit Gaussian distribution $p \sim N(0, 1)$, the learned style feature is regularized to suppress contained structure information through reconstruction.

The KL-divergence loss limits the distribution range and capacity of the style feature. By fusing inferred style code z with encoded structure representation, sufficient structure information can be obtained from prior through multi-level skip connection. Extra structure encoded in z incurs penalty of the likelihood $p(x|y, z)$ during training with no new information captured. In this way, E^{style} is discouraged from learning structure information that is provided by E^{struct} during training. To better reconstruct the original image, E^{style} is enforced to learn structure-invariant style information.

Reconstruction Loss The second term \mathcal{L}_{rec} in Eq. (2) refers to the reconstruction loss in the auto-encoder frame-

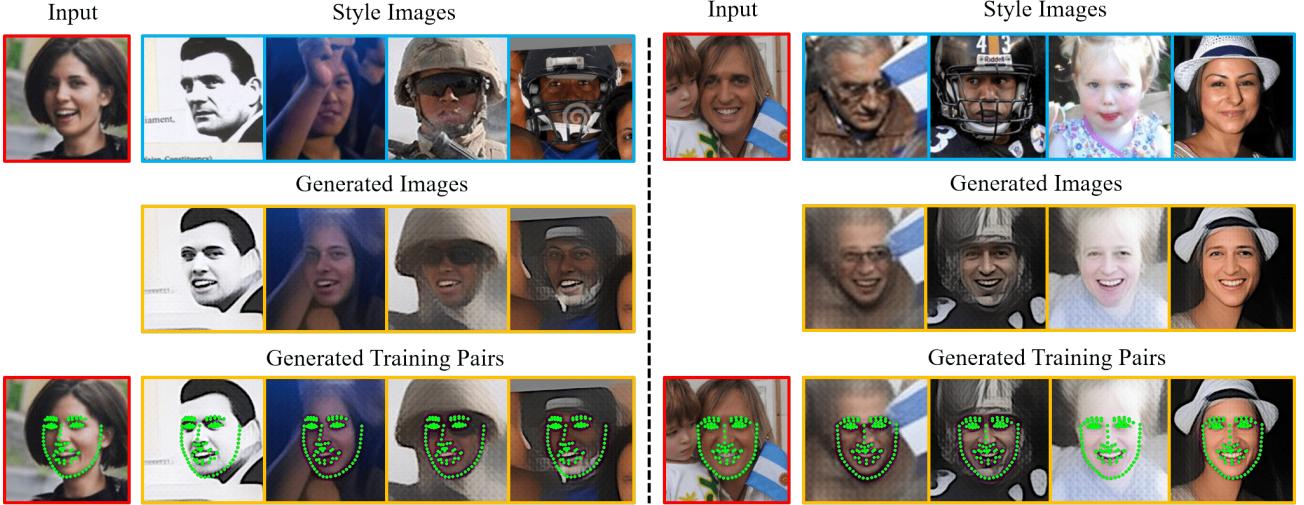


Figure 3: Visualization of style translation. Given the input images in red 4 different styles are provided to perform translation towards input structure. The synthetic images along with input original landmarks are provided to demonstrate the strong coherence of structure.

work. As widely discussed [81, 30], basic pixel-wise L_1 or L_2 loss cannot model rich information within images well. We instead adopt perceptual loss to capture style information and better visual quality. \mathcal{L}_{rec} is formulated as

$$\mathcal{L}_{rec}(x, \theta, \phi) = \sum_l \|\Phi_l(x) - \Phi_l(G^{full}(x, y))\|_2^2, \quad (3)$$

where we use VGG-19 network Φ structure that measures perceptual quality. l indexes the layer of network Φ .

Since the style definition could be complicated, E^{style} here encodes semantics of the style signal that simulates different types of degradation. It does not have to maintain fine-grained visual details. Besides, to reserve the strong prior on structure information encoded from landmarks y , skip connection between E^{struct} and D^{render} is established to avoid landmark inaccuracy through style translation.

In this design, the model is capable of learning complementary representation of facial geometry and image style.

3.2. Augmenting Training via Style Translation

Disentanglement of structure and style forms a solid foundation for diverse stylized face images under invariant structure prior.

Given a dataset X that contains n face images with landmarks annotation, each face image $x_i (1 \leq i \leq n)$ within the dataset has its explicit structure denoted by landmark y_i , as well as an implicit style code z_i depicted and embedded by E^{style} . To perform style translation between two images x_i and x_j , we pass their latent style and structure code embedded by E^{style} and E^{struct} to D^{render} . To put the style of image x_j on x_i 's structure, the stylized synthetic image

is denoted as

$$x_{ij} = D^{render}(E^{style}(x_j, y_i), E^{struct}(y_i)). \quad (4)$$

As illustrated in Fig. 2, the first stage of our framework is to train the disentangling components. In the second phase, by augmenting and rendering a given sample x in the original dataset X with styles from random k other faces, we produce $k \times n$ “stylized” synthetic face images with respective annotated landmarks. These samples are then fed into training of facial landmark detectors together with the original dataset. Visualization of style translation results is provided in Fig. 3. The input facial geometry is maintained under severe style variation, indicating its potential at augmenting training of facial landmark detectors.

Albeit with cohesive structure, the decoder generally does not re-render perfect-quality images, since the complexity of plentiful style information has been diminished to a parametric Gaussian distribution, confined by its capacity. Also, as discussed before, each face image x_i has its own style. Theoretically, the renderer could synthesize n^2 images by rendering each available landmark with any other images' style. To understand how the quantity of stylized synthetic samples helps improve the facial landmark detectors, we analyze the effect of our design in following experiments and ablation study.

4. Experiments

4.1. Datasets

WFLW [69] dataset is a challenging one, which contains 7,500 faces for training and 2,500 faces for testing, based

Metric	Method	Fullset	Pose	Expression	Illumination	Make-Up	Occlusion	Blur
Mean Error (%)	CFSS [82]	9.07	21.36	10.09	8.30	8.74	11.76	9.96
	DVLN [70]	6.08	11.54	6.78	5.73	5.98	7.33	6.88
	LAB [69]	5.27	10.24	5.51	5.23	5.15	6.79	6.32
	SAN [15]	5.22	10.39	5.71	5.19	5.49	6.83	5.80
	WING [19]	5.11	8.75	5.36	4.93	5.41	6.37	5.81
	Res-18	6.09	10.76	6.97	5.83	6.19	7.15	6.67
	Ours w. Res-18	5.25	9.10	5.83	4.93	5.47	6.26	5.86
	Ours w. LAB	4.76	8.21	5.14	4.51	5.00	5.76	5.43
Failure Rate (%)	CFSS [82]	20.56	66.26	23.25	17.34	21.84	32.88	23.67
	DVLN [70]	10.84	46.93	11.15	7.31	11.65	16.30	13.71
	LAB [69]	7.56	28.83	6.37	6.73	7.77	13.72	10.74
	SAN [15]	6.32	27.91	7.01	4.87	6.31	11.28	6.60
	WING [19]	6.00	22.70	4.78	4.30	7.77	12.50	7.76
	Res-18	10.92	43.87	13.38	7.31	11.17	16.30	11.90
	Ours w. Res-18	7.44	32.52	8.60	4.30	8.25	12.77	9.06
	Ours w. LAB	5.24	20.86	4.78	3.72	6.31	9.51	7.24
AUC @0.1	CFSS [82]	0.3659	0.0632	0.3157	0.3854	0.3691	0.2688	0.3037
	DVLN [70]	0.4551	0.1474	0.3889	0.4743	0.4494	0.3794	0.3973
	LAB [69]	0.5323	0.2345	0.4951	0.5433	0.5394	0.4490	0.4630
	SAN [15]	0.5355	0.2355	0.4620	0.5552	0.5222	0.4560	0.4932
	WING [19]	0.5504	0.3100	0.4959	0.5408	0.5582	0.4885	0.4918
	Res-18	0.4385	0.1527	0.3718	0.4559	0.4366	0.3655	0.3931
	Ours w. Res-18	0.5034	0.2294	0.4534	0.5252	0.4849	0.4318	0.4532
	Ours w. LAB	0.5460	0.2764	0.5098	0.5660	0.5349	0.4700	0.4923
	Ours w. SAN	0.5913	0.3109	0.5490	0.6089	0.5812	0.5164	0.5513

Table 1: Evaluation of our approach on WFLW dataset. Top-2 results are highlighted in bold font.

on WIDER Face [74] with 98 manually annotated landmarks [69]. The dataset is partitioned into 6 subsets according to challenging attribute annotation of large pose, expression, illumination, makeup, occlusion, and blur.

300W [56] provides multiple face datasets including LFWP [4], AFW [54], HELEN [41], XM2VTS [49], and IBUG with 68 automatically-annotated landmarks. Following the protocol used in [55], 3,148 training images and 689 testing images are used. The testing images include two subsets, where 554 test samples from LFWP and HELEN form the common subset and 135 images from IBUG constitute the challenging subset.

AFLW [36] dataset is widely used for benchmarking facial landmark localization. It contains 24,386 in-the-wild faces with a wide range of yaw, pitch and roll angles ($[-120^\circ, 120^\circ]$ for yaw, $[-90^\circ, 90^\circ]$ for pitch and roll). Following the widely-adopted protocol [82, 83], the AFLW-full dataset has 20,000 images for training and 4,386 for testing. It is originally annotated with 19 sparse facial landmarks. To provide a better benchmark for evaluating pose variation and allow cross-dataset evaluation, we re-annotate it with 68 facial landmarks, which follow the common standard in 300W [56, 58]. Based on the new 68-point annotation, we conduct more precise evaluation. Cross-dataset evaluation

is also provided among existing datasets [4, 54, 41].

COFW dataset [7] contains 1,345 images for training and 507 images for testing, focusing on occlusion. The whole dataset is originally annotated with 29 landmarks and has been re-annotated with 68 landmarks in [22] to allow cross-dataset evaluation. We utilize 68 annotated landmarks provided by [22] to conduct comparison with other approaches.

4.2. Experimental Setting

Evaluation Metrics We evaluate performance of facial landmark detection using normalized landmarks mean error and Cumulative Errors Distribution (CED) curve. For the 300W dataset, we normalize the error using inter-pupil distance. In Table 2, we also report the NME using inter-ocular distance to compare with algorithms of [15, 31, 71, 37], which also use it as the normalizing factor. For other datasets, we follow the protocol used in [56, 63] and apply inter-ocular distance for normalization.

Implementation Details Before training, all images are cropped and resized to 256×256 using provided bounding boxes. For the detailed conditional variational auto-encoder network structures, we use a two-branch encoder-decoder structure as shown in Fig. 2. We use 6 residual encoder blocks for downsampling the input feature maps,

where batch normalization is removed for better synthetic results. The facial landmark detector backbone is substitutable and different detectors are usable to achieve improvement, which we will discuss later.

For training of the disentangling step, we use Adam [34] with an initial learning rate of 0.01, which descends linearly to 0.0001 with no augmentation. For training of detectors, we first augment each landmark map with k random styles sampled from other face images. The number is set to 8 if not specially mentioned in experiments. For the detector architecture, a simple baseline network based on ResNet-18 [24] is chosen by changing the output dimension of the last FC layers to $\text{landmark} \times 2$ to demonstrate the increase brought by style translation. To compare with state-of-the-arts and further validate the effectiveness of our approach, we replace our baseline model with similar structures proposed in [69, 15], with the same affine augmentation.

4.3. Comparison with State-of-the-arts

WFLW We evaluate our approach on WFLW [69] dataset. WFLW is a recently proposed challenging dataset with images from in-the-wild environment. We compare algorithm in terms of NME(%), Failure Rate(%) and AUC(@0.1) following protocols used in [69].

The Res-18 baseline receives strong enhancement using synthetic images. To further verify the effectiveness and generality of using style information, we replace the network by two strong baselines [15, 69] and report the result in Table 1. The light-weight Res-18 is improved by 13.8%. By utilizing a stronger baseline, our model achieves 4.39% NME under style-augmented training, outperforms state-of-the-art entries by a large margin. In particular, for the strong baselines, our method also brings 15.9% improvement to SAN [15] model, and 9% boost to LAB [69] from 5.27% NME to 4.76%. The elevation is also determined by the model capacity.

300W In Table 2, we report different facial landmark detector performance (in terms of normalized mean error) on 300W dataset. The baseline network follows Res-18 structure. With additional “style-augmented” synthetic training samples, our model based on a simple backbone outperforms previous state-of-the-art methods. We also report results of models that are trained on original data, which reflect the performance gain brought by our approach.

Similarly, we replace the baseline model with a state-of-the-art method [15]. Following the same setting, this baseline is also much elevated. Note that the 4-stack LAB [69] and SAN [15] are open-source frameworks. We train the models from scratch, which perform less well than those reported in their original papers. However, our model still yields 1.8% and 3.1% improvement on LAB and SAN respectively, which manifest the consistent benefit when using the “style-augmented” strategy.

Method	Common Subset	Challenging Subset	Fullset
Inter-pupil Normalization			
SDM [73]	5.57	15.40	7.52
CFAN [77]	5.50	16.78	7.69
ESR [8]	5.28	17.00	7.58
LBF [55]	4.95	11.98	6.32
CFSS [82]	4.73	9.98	5.76
TCDCN [80]	4.80	8.60	5.54
RCN [27]	4.67	8.44	5.41
3DDFA [84]	6.15	10.59	7.01
SeqMT [26]	4.84	9.93	5.74
RAR [72]	4.12	8.35	4.94
TSR [45]	4.36	7.56	4.99
DCFE [65]	3.83	7.54	4.55
LAB [69]	4.20	7.41	4.92
Res-18	4.53	8.41	5.30
Ours w LAB	4.23	7.32	4.83
Ours w Res-18	3.98	7.21	4.54
Inter-ocular Normalization			
PIFA [31]	5.43	9.88	6.30
RDR [71]	5.03	8.95	5.80
PCD-CNN [37]	3.67	7.62	4.44
SAN [15]	3.34	6.60	3.98
Ours w SAN	3.21	6.49	3.86

Table 2: Normalized mean error (%) on 300W common, challenging subset and the full set.

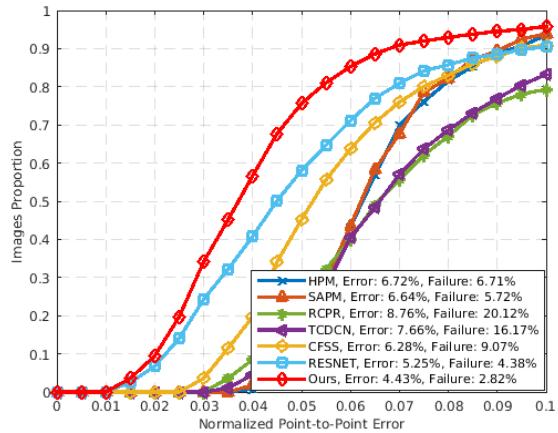


Figure 4: Cumulative error distribution curve on COFW 68-point test set.

Cross-dataset Evaluation on COFW To comprehensively evaluate the robustness of our method towards occlusion, COFW-68 is also utilized for cross-dataset evaluation. We perform comparison against several state-of-the-art methods in Fig. 4. Our model performs the best with 4.43% mean error and 2.82% failure rate, which indicates high robustness to occlusion due to our proper utilization of style translation.

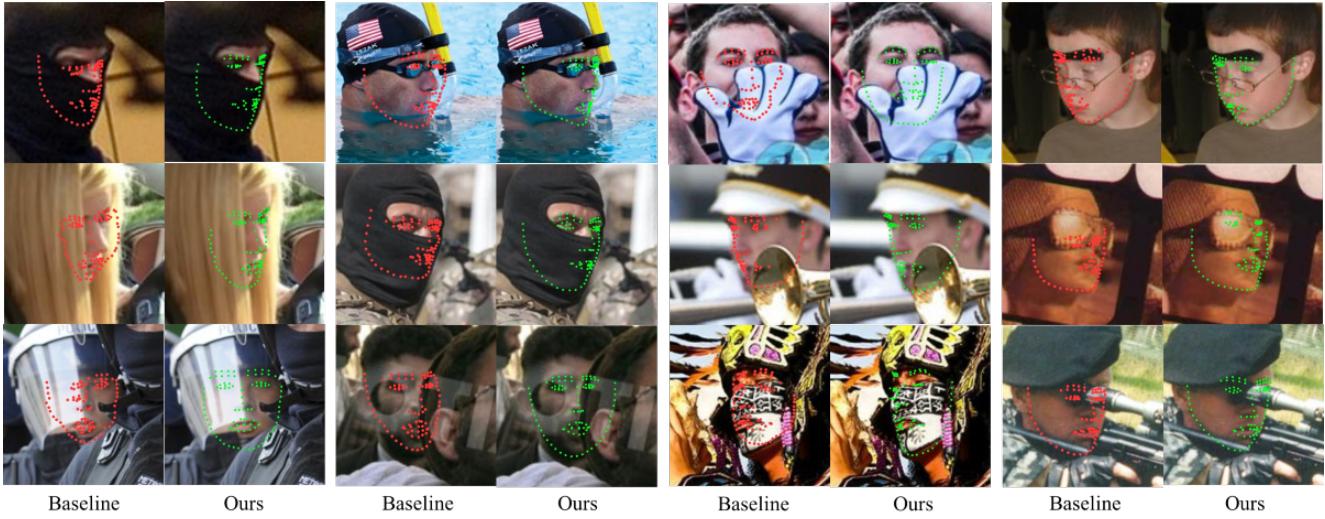


Figure 5: Visual comparison on WFLW test set between the original baseline model and the boosted framework via style translation.

Method	NME(%)		AUC@ 0.1	FR(%)
	Frontal	Full	Full	Full
LAB [69]	2.23	7.15	0.39	11.28
SAN [15]	2.01	6.94	0.44	10.43
Res-18	2.30	7.23	0.37	11.89
Ours w. Res-18	2.20	7.17	0.38	11.91
Ours w. LAB	2.10	7.06	0.42	10.01
Ours w. SAN	1.86	6.01	0.58	9.70

Table 3: Normalized mean error (%) on re-annotated 68-pt AFLW frontal subset and the full set.

AFLW We further evaluate our algorithm on the AFLW [36] dataset following the AFLW Full protocol. AFLW is also challenging for its large pose variation. It is originally annotated with 19 facial landmarks, which are relatively sparse. To make it more useful, we richen the dataset by re-annotating it with 68-point facial landmarks. This new set of data is also publicly available.

We compare our approach with several models in Table 3, by re-implementing their algorithms on the new dataset along with our style-augmented samples. Exploiting style information also boosts landmark detectors with a large-scale training set (25,000 images in AFLW). Interestingly, our method improves SAN baseline in terms of NME on Full set from 6.94% to 6.01%, which indicates that augmenting in style level brings promising improvement on solving large pose variation. The visual comparison in Fig. 5 shows hidden face part is better modeled with our strategy.

4.4. Ablation Study

4.4.1 Improvement on Limited Data

Disentanglement of style and structure is the key that influences quality of style-augmented samples. We evaluate the

Dataset	PCT (%)	NME (%)		
		Res-18	w Ours	Improved
300W	10	13.72	7.86	+42.71%
	20	9.66	6.07	+37.16%
	30	8.9	5.86	+34.16%
	40	8.86	5.29	+40.29%
	50	7.96	5.23	+34.30%
	60	7.89	5.18	+34.35%
	70	7.02	5.04	+28.21%
	80	6.66	4.82	+27.63%
	90	6.58	4.69	+28.72%
WFLW	10	22.09	10.81	+51.06%
	20	16.04	8.98	+44.01%
	30	13.91	8.24	+40.76%
	40	12.19	8.03	+34.13%
	50	11.78	7.75	+34.21%
	60	10.41	7.31	+29.78%
	70	9.87	7.29	+26.14%
	80	9.66	7.25	+24.95%
	90	9.04	7.19	+20.46%

Table 4: Normalized mean error (%) on 300W common and WFLW datasets when the training images are split into 10 folds. Each row represents NME on test set when the model is trained using a percentage (PCT%) of the training set. The landmark detector backbone is Res-18.

completeness of disentanglement especially when the training samples are limited. To evaluate the performance and relative gain of our approach when training data is limited. The training set is split into 10 subsets and respectively we evaluate our model on different portions of training data. Note that for different portions, we train the model from scratch with no extra data used. The quantitative result is reported in Tables 4 and 5.

In Table 4, a light baseline network Res-18 is used to

Dataset	PCT (%)	NME (%)		
		SAN	w Ours	Improved
300W	10	84.33	4.27	+94.94%
	20	5.08	3.85	+24.21%
	30	4.05	3.65	+9.88%
	40	3.8	3.49	+8.16%
	50	3.6	3.39	+5.83%
	60	3.54	3.32	+6.21%
	70	3.48	3.29	+5.46%
	80	3.39	3.21	+5.31%
	90	3.38	3.19	+5.62%
	10	9.16	7.2	+21.40%
WFLW	20	7.41	6	+19.03%
	30	6.73	5.48	+18.57%
	40	6.26	5.21	+16.77%
	50	5.95	4.98	+16.30%
	60	5.72	4.84	+15.38%
	70	5.5	4.69	+14.73%
	80	5.43	4.63	+14.73%
	90	5.23	4.6	+12.05%

Table 5: Normalized mean error (%) on 300W common and WFLW datasets when using different percentages of the training set, with the same protocol as in Table 4 on a stronger baseline. The baseline network here follows SAN [15] structure.

show the relative improvement on different training samples. Style-augmented synthetic images improve detectors' performance by a large margin, while the improvement is even larger when the number of training images is quite small. In Table 5, a stronger baseline SAN [15] is chosen. Surprisingly, the baseline easily reaches state-of-the-art performance using only 50% labeled images, compared to former methods provided in Table 1.

Besides, Fig. 6 provides an intuitive visualization of the resulting generated faces when part of the data is used. Each column contains output that is rendered from the input structure and given style, when using a portion of face image data. It shows when the data is limited, our separation component tends to capture weak style information, such as color and lighting. Given more data as examples, the style becomes complex and captures detailed texture and degradation, like occlusion.

The results verify that even using limited labeled images, our design is capable of disentangling style information and keeps improve those baseline methods that are already very strong.

4.4.2 Estimating the Upper-bound

As discussed before, our method conceptually and empirically augments training with n^2 synthetic samples. By aug-

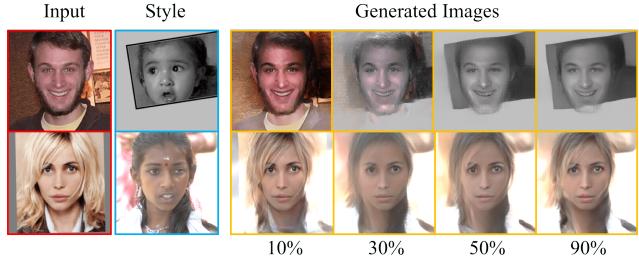


Figure 6: Results of style translation using different numbers of data. The left 2 images are the input, with 2 different reference styles. The percentage refers to how much data is used to train the disentangle module.

Number	0	2	4	8	16	32
NME (%)	6.22	5.89	5.54	5.31	5.29	5.34

Table 6: Normalized mean error (%) on WFLW test set using different numbers of style translation.

menting each face image with k random styles, the training set could be very large and slows down convergence. In this section, we experiment with choosing the style augmenting factor k and test the upper bound of style translation. We evaluate our method by adding the number of random sampled styles k of each annotated landmarks on a ResNet-50 baseline.

The result is reported in Table 6. By adding a number of augmented styles, the model continue gaining improvement. However, when $k \geq 8$, the performance grow slows down. It begins to decrease if k reaches 32. The reason is that due to the quantity imbalance between real and synthetic faces, a very large k makes the model overfit to synthetic image texture when the generated image quantity is large.

5. Conclusion and Future Work

In this paper, we have analyzed the well-studied facial landmark detection problem from a new perspective of implicit style and environmental factor separation and utilization. Our approach exploits the disentanglement representation of facial geometry and unconstrained style to provide synthetic faces via style translation, further boosting quality of facial landmarks. Extensive experimental results manifest its effectiveness and superiority.

We also note that utilizing synthetic data for more high-level vision tasks still remains an open problem, mainly due to the large domain gap between generated and real images. In our future work, we plan to model style in a more realistic way by taking into account the detailed degradation types and visual quality. We also plan to generalize our structure to other vision tasks.

References

- [1] Anthreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks, 2018. [12](#)
- [2] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *CVPR*, 2018. [2](#)
- [3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Devon Hjelm, and Aaron Courville. Mutual information neural estimation. In *ICML*, 2018. [2](#)
- [4] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013. [5, 13](#)
- [5] Adrian Bulat and Georgios Tzimiropoulos. Convolutional aggregation of local evidence for large pose face alignment. 2016. [2](#)
- [6] Adrian Bulat and Georgios 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. [1](#)
- [7] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *CVPR*, 2013. [1, 2, 5](#)
- [8] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *IJCV*, 2014. [2, 6](#)
- [9] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *ECCV*, 2014. [2](#)
- [10] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016. [1, 2](#)
- [11] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2001. [2](#)
- [12] David Cristinacce and Timothy F Cootes. Feature detection and tracking with constrained local models. In *Bmvc*, 2006. [2](#)
- [13] Jiankang Deng, George Trigeorgis, Yuxiang Zhou, and Stefanos Zafeiriou. Joint multi-view face alignment in the wild. *arXiv preprint arXiv:1708.06023*, 2017. [2](#)
- [14] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*. [2](#)
- [15] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *CVPR*, 2018. [1, 2, 5, 6, 7, 8](#)
- [16] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018. [1, 2](#)
- [17] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. [1](#)
- [18] Zhen-Hua Feng, Guosheng Hu, Josef Kittler, William Christmas, and Xiao-Jun Wu. Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting. *IEEE Transactions on Image Processing*, 24(11):3425–3440, 2015. [2](#)
- [19] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *CVPR*, 2018. [1, 5](#)
- [20] Peter Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200, 1991. [2](#)
- [21] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. [1, 2, 13](#)
- [22] Golnaz Ghiasi and Charless C Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *CVPR*, 2014. [5, 13](#)
- [23] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. [13](#)
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016. [6](#)
- [25] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. [1](#)
- [26] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *CVPR*, 2018. [2, 6](#)
- [27] Sina Honari, Jason Yosinski, Pascal Vincent, and Christopher Pal. Recombinator networks: Learning coarse-to-fine feature aggregation. In *CVPR*, 2016. [6](#)
- [28] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. [2](#)
- [29] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. [2](#)
- [30] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. [1, 4](#)
- [31] Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. Pose-invariant face alignment with a single cnn. In *CVPR*, 2017. [1, 5, 6](#)
- [32] Fatih Kahraman, Muhittin Gokmen, Sune Darkner, and Rasmus Larsen. An active illumination and appearance (aia) model for face alignment. In *CVPR*, 2007. [2](#)
- [33] Muhammad Haris Khan, John McDonagh, and Georgios Tzimiropoulos. Synergy between face alignment and tracking via discriminative global consensus optimization. In *ICCV*, 2017. [1](#)
- [34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [6](#)

- [35] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2
- [36] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCV workshops*. IEEE, 2011. 1, 2, 5, 7, 13
- [37] Amit Kumar and Rama Chellappa. Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. In *CVPR*, 2018. 1, 5, 6
- [38] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *ECCV*, 2016. 2
- [39] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *CVPR*, 2017. 2
- [40] Quoc V Le, Will Y Zou, Serena Y Yeung, and Andrew Y Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. 2011. 2
- [41] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *ECCV*, 2012. 5, 13
- [42] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 1, 2
- [43] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017. 2
- [44] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *ICCV*, 2017. 1
- [45] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *CVPR*, 2017. 2, 6
- [46] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NIPS*, 2017. 2
- [47] Iain Matthews and Simon Baker. Active appearance models revisited. *IJCV*. 2
- [48] Daniel Merget, Matthias Rock, and Gerhard Rigoll. Robust facial landmark detection via a fully-convolutional local-global context network. In *CVPR*, 2018. 2
- [49] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966, 1999. 5, 13
- [50] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. Direct shape regression networks for end-to-end face alignment. In *CVPR*. 2
- [51] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 2
- [52] Xi Peng, Rogerio S Feris, Xiaoyu Wang, and DN Metaxas. A recurrent encoder-decoder for sequential face alignment. In *ECCV*. 2
- [53] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *CVPR*, 2018. 2
- [54] Deva Ramanan and Xiangxin Zhu. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012. 5, 13
- [55] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *CVPR*, 2014. 5, 6
- [56] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV Workshops*, 2013. 1, 2, 5, 13
- [57] Jason Saragih and Roland Goecke. A nonlinear discriminative approach to aam fitting. In *ICCV*, 2007. 2
- [58] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kosaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *CVPR Workshops*, 2015. 5
- [59] David Stavens and Sebastian Thrun. Unsupervised learning of invariant features using video. In *CVPR*, 2010. 2
- [60] Graham W Taylor, Rob Fergus, Yann LeCun, and Christoph Bregler. Convolutional learning of spatio-temporal features. In *ECCV*, 2010. 2
- [61] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016. 1
- [62] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017. 1
- [63] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *CVPR*, 2016. 2, 5
- [64] Oncel Tuzel, Tim K Marks, and Salil Tambe. Robust face alignment using a mixture of invariant experts. In *ECCV*, 2016. 2
- [65] Roberto Valle, Jose M Buenaposada, Antonio Valdes, and Luis Baumela. A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment. In *ECCV*, 2018. 6
- [66] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *CVPR*, 2015. 2
- [67] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 2
- [68] Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002. 2
- [69] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, 2018. 1, 2, 4, 5, 6, 7, 12
- [70] Wayne Wu and Shuo Yang. Leveraging intra and inter-dataset variations for robust face alignment. In *CVPR*, 2017. 5
- [71] Shengtao Xiao, Jiashi Feng, Luoqi Liu, Xuecheng Nie, Wei Wang, Shuicheng Yan, and Ashraf Kassim. Recurrent 3d-2d dual learning for large-pose facial landmark detection. In *ICCV*, 2017. 5, 6

- [72] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shucheng Yan, and Ashraf Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *ECCV*, 2016. [2](#), [6](#)
- [73] Xuehan Xiong and Fernando De la Torre. Global supervised descent method. In *CVPR*, 2015. [2](#), [6](#)
- [74] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaou Tang. Wider face: A face detection benchmark. In *CVPR*, 2016. [5](#)
- [75] Stefanos Zafeiriou, George Trigeorgis, Grigoris Chrysos, Jiankang Deng, and Jie Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *CVPR Workshops*, 2017. [2](#)
- [76] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. [2](#)
- [77] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *ECCV*, 2014. [6](#)
- [78] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. [2](#)
- [79] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, 2017. [2](#)
- [80] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 2016. [2](#), [6](#)
- [81] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. [4](#)
- [82] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaou Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, 2015. [2](#), [5](#), [6](#)
- [83] Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaou Tang. Unconstrained face alignment via cascaded compositional learning. In *CVPR*, 2016. [2](#), [5](#)
- [84] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016. [6](#)
- [85] Will Zou, Shenghuo Zhu, Kai Yu, and Andrew Y Ng. Deep learning of invariant features via simulated fixations in video. In *NIPS*, 2012. [2](#)

Appendix

The content of our supplementary material is organized as follows.

1. More ablation studies and detailed analysis of components in our framework.
2. Additional discussion about related directions.
3. Details of our annotated AFLW-68 dataset and some representative visualized samples.

S1. More Ablation Studies

In this section, we provide additional analysis about each design in our framework to facilitate understanding of our structure. Two key loss terms in our framework are studied to give insights into their respective roles. Qualitative visualization and quantitative results are reported for a comprehensive comparison.

KL divergence loss and perceptual loss, are incorporated into our framework during the disentangled learning procedure. Fig. 7 shows their respective effect on style translation via visual comparisons of several incomplete variants. Through visual observations, their roles could be inferred intuitively. The perceptual loss, as discussed, is designed to capture better style information and visual quality. Thus, removing this term leads to “over-smoothness” and poor diversity on synthetic images. Removing KL divergence term shows severe structure distortion on translated results, which indicates that KL divergence loss plays a key role on disentangling structure and style information.

Quantitative results of each variants are also reported in Table. 7. The normalized mean error(NME) is evaluated on WFLW [69] test set when the model is trained on style augmented dataset using each variant. We observe that NME will increase if any loss function is removed. In particular, the detector performance drops significantly lower than the baseline if L_{KL} is removed. Both the qualitative and quantitative result interprets the role of each component, indicating their essentialness in our framework.

Model	Baseline	wo KL divergence	wo Perceptual	Full
NME(%)	8.49	9.08	8.34	7.98

Table 7: Quantitative ablative results. Normalized mean error (%) on WFLW test set using different variants of our framework.

S2. Additional Discussion

In this section, we provide more discussion on our approach along with our analysis towards some existing alternatives.

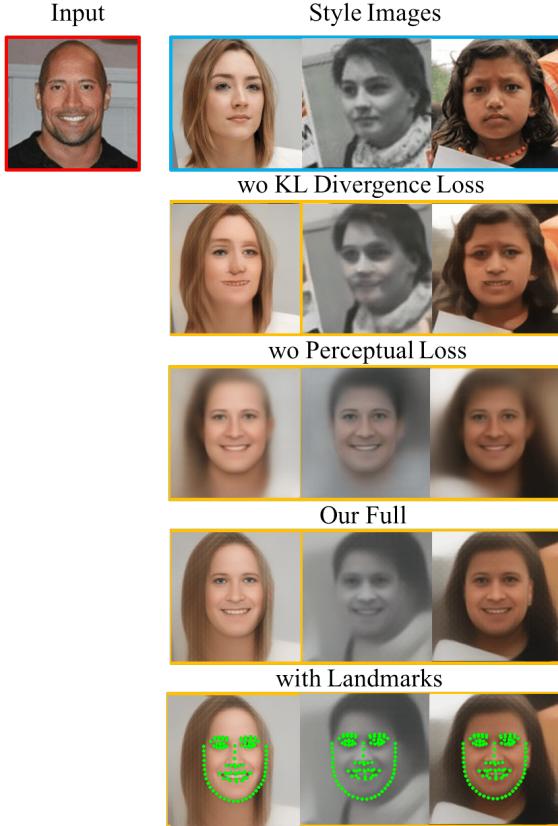


Figure 7: Qualitative analysis of each component in our framework. Given input images in red, 3 different styles are provided to perform translation towards input structure. 3 incomplete variants of our framework are used to show the functionality of each component.

S2.1 Comparison with GAN-based approaches

Generative adversarial network (GAN) and its applications are widely studied these days, using GAN-synthetic data to aid training, has also been explored along this line. Some works [1] have utilized GANs to perform data augmentation. However, its effect still remains questionable especially on high-level vision challenges. For instance, in our task, face images need to be labeled with accurate landmarks. Existing generative models are incapable of handling these tasks with fine-grained annotations, e.g. semantic segmentation, constrained by its limited generalizability. We choose to escape the difficulties of GAN training, starting from a new perspective of internal representation. With decent representation of separating style and structure, different interactions within a face image can be simulated by re-rendering from existing style and structure code. In other words, our choice depends upon fully exploiting available information by mixing them, instead of creating new information and visually perfect results via adversarial learning procedure. However, if two codes of structure and style are

factored well, advances on high fidelity images synthesis could theoretically bring more gains based on our framework.

S2.2 Comparison with Style Transfer

Our method is motivated by advances in style transfer. A common doubt could be why not directly conducting style transfer as a augmentation or how basic style transfer could help training. As discussed, our definition of style includes environments and degradation that prevent the model from recognizing while content refers to facial geometry. Applying “vanilla” style transfer would leads to structural distortion on stylized images, as illustrated in Fig. 8. Our definition of “style” helps preserve structure on synthetic images. Besides, synthetic images using style transfer have a large domain gap with real-world face images. Simply augmenting training with these samples would instead hurt model’s localization ability on real images.



Figure 8: Visual comparison with style transfer approach. For the style transfer algorithm, we use [21]. Our results are more realistic than stylized images, with better structure coherence.

Database	Environment	Number
Multi-PIE [23]	Controlled	750000
XM2VTS [49]		2360
LFPW [4]		1035
HELEN [41]	In-the-wild	2330
AFW [54]		468
IBUG		135
COFW-68 [22]		507
AFLW-68(Ours)		25993

Table 8: Widely-used 68-pt facial landmark datasets. Dataset names their the environment and number are reported.

S3. Details of AFLW 68-point dataset

We propose a new facial landmark dataset based on AFLW [36], to facilitate benchmarking on large pose performance. To allow a more precise evaluation and cross-dataset comparison, we follow the widely-used Multi-



Figure 9: Sampled annotated images in the proposed AFLW 68-point dataset, including in-the-wild faces under large pose variations

PIE [23] and 300W [56] 68-point protocol. Annotated samples are provided at Fig. 9, which contains extreme pose variations.