FreeEnricher: Enriching Face Landmarks without Additional Cost

Yangyu Huang, Xi Chen, Jongyoo Kim*, Hao Yang, Chong Li, Jiaolong Yang, Dong Chen

Microsoft Research Asia
{yanghuan, xichen6, jongk, haya, chol, jiaoyan, doch}@microsoft.com,

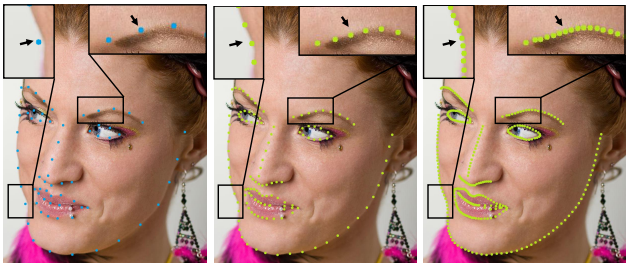
Abstract

Recent years have witnessed significant growth of face alignment. Though dense facial landmark is highly demanded in various scenarios, e.g., cosmetic medicine and facial beautification, most works only consider sparse face alignment. To address this problem, we present a framework that can enrich landmark density by existing sparse landmark datasets, e.g., 300W with 68 points and WFLW with 98 points. Firstly, we observe that the local patches along each semantic contour are highly similar in appearance. Then, we propose a weakly-supervised idea of learning the refinement ability on original sparse landmarks and adapting this ability to enriched dense landmarks. Meanwhile, several operators are devised and organized together to implement the idea. Finally, the trained model is applied as a plug-and-play module to the existing face alignment networks. To evaluate our method, we manually label the dense landmarks on 300W testset. Our method yields state-of-the-art accuracy not only in newly-constructed dense 300W testset but also in the original sparse 300W and WFLW testsets without additional cost.

Introduction

Face alignment provides semantically consistent facial landmarks in given face photos, which plays a critical role in several face-related vision tasks, such as face recognition (Masi et al. 2018), face parsing (Luo, Wang, and Tang 2012), and face reconstruction (Jiang et al. 2005). Due to its importance, great efforts are devoted to the development of algorithms as well as the construction of datasets. Recently, remarkable alignment accuracy could be achieved through deep neural networks (Wu et al. 2018; Wang, Bo, and Fuxin 2019; Kumar et al. 2020). To successfully train such deep models, a large number of face alignment datasets have been published, e.g., COFW (Burgos-Artizzu, Perona, and Dollár 2013), 300W (Sagonas et al. 2013) and WFLW (Wu et al. 2018).

Despite of enormous growth in face alignment, there have been very few researches handling dense facial landmarks. In regard to real-world applications, dense landmarks are highly demanded in a lot of scenarios, for instance, measuring facial metrics in a medical application and delicate face editing and reenactment. The fundamental bottleneck is the substantial cost of labeling dense facial landmarks. To tackle this problem, we propose a novel framework for enriching landmarks[1] to benefit both academic and industrial scenarios.



(a) ADNet (Baseline) (b) ADNet-FE2 (Ours) (c) ADNet-FE5 (Ours)

Figure 1: Comparison of baseline method and our FreeEnricher method. (a) is the detected sparse landmarks by baseline ADNet (Huang et al. 2021); (b) and (c) are the enriched dense landmarks by our FreeEnricher with ADNet. The definitions of ADNet-FE2 and ADNet-FE5 could refer to Table 1. Compared to the baseline method, FreeEnricher could enrich landmarks to specific densities with higher accuracy. Meanwhile, FreeEnricher merely relies on the original 68-points 300W training dataset without both additional annotation and extra inference time.

In the literature, a 3D face model is introduced to handle dense face alignment, where the 3D model is fit to images to map each pixel to the 3D surface of the face template (Zhu et al. 2016; Liu et al. 2017, 2016; Cootes et al. 1995). By defining the landmark positions on the 3D face template, they could provide dense landmarks. However, these approaches suffer from the limited capacity of the 3D face model and the 2D to 3D ambiguity of the semantic landmark definition, which results in inaccurate dense landmarks and inconsistency with 2D features.

To mitigate these issues, we propose a weakly-supervised approach, FreeEnricher, to freely enrich the pre-annotated

*Corresponding author

[1]We denote *enriching landmarks* by increasing landmark points along the landmark contours.

landmarks without any additional annotation tasks. Rather than relying on the 3D face model, we focus on reliable 2D features and informative landmark contours. Specifically, we firstly observe that the local image patches are highly similar in appearance along each semantic facial component, e.g., eye, lip, and face contour. Activated by it, we devise an idea of learning refinement ability on original sparse landmarks and transferring this ability to initialized dense landmarks. Then a unified framework is designed to implement the idea by organizing several operators together. Finally, FreeEnricher is plugged into the existing face alignment networks, e.g. state-of-the-art ADNet. The new network can generate dense facial landmarks while yielding state-of-the-art accuracy. Fig. 1 compares the results of ADNet (Huang et al. 2021) and our ADNet-FE2/5. While ADNet can only predict the sparse landmarks as shown in (a), ADNet-FE2 and ADNet-FE5 infer highly dense and more accurate points as shown in (b) and (c).

Our contributions can be summarized as follows:

- We propose a novel method, FreeEnricher, which should be the first method that can freely enrich the facial landmark density without additional cost with respect to both manual annotation and inference time.
- We devise a weakly-supervised framework of FreeEnricher to transfer the refinement ability from original sparse landmarks to enriched dense landmarks. Based on it, FreeEnricher can enrich landmarks to arbitrary density and be plug-and-play to existing alignment networks.
- We release the enriched 300W with both preprocessed training set and annotated testing set, which would leverage the future work in the face alignment field.
- Our network, plugging FreeEnricher to ADNet, achieves not only accurate dense landmarks in the newly-constructed **enriched** 300W but also the state-of-the-art accuracy in the **original** 300W and WFLW testing sets.

Related Work

Face alignment has been studied widely and produced fruitful outputs. Early methods such as AAMs (Cootes, Edwards, and Taylor 2001; Saragih and Goecke 2007; Matthews and Baker 2004) and ASMs (Cootes and Taylor 1992; Cootes et al. 1995; Milborrow and Nicolls 2008) are based on a generative model. Later, the direct regression method became popular due to its simplicity and high accuracy, which can be generally categorized into coordinate and heatmap regression approaches.

Coordinate Regression. In regression-based methods (Sun, Wang, and Tang 2013; Toshev and Szegedy 2014; Trigeorgis et al. 2016; Lv et al. 2017; Zhou et al. 2013; Zhang et al. 2014a), landmark coordinates are directly regressed by the CNN network. To improve the localization ability of landmarks, a sort of methods (Zhang et al. 2014a; Trigeorgis et al. 2016; Sun, Wang, and Tang 2013) apply coarse-to-fine strategy to refine the position. Additionally, Smooth L1 (Girshick 2015) and Wing loss (Feng et al. 2018) optimize the loss function by reducing the weight of outlier. Recently, methods of extra information prediction, such as uncertainties (Kumar et al. 2020; Gundavarapu et al. 2019), push regression-based methods to a new climax.

Heatmap Regression. To take full advantage of CNN networks, 2D heatmaps are utilized as intermediate outputs, then coordinates are inferred using non-maximum suppression or soft-argmax (Deng et al. 2019; Dong et al. 2018; Newell, Yang, and Deng 2016; Wei et al. 2016). Several studies propose novel CNN architectures such as HRNet (Wang et al. 2020), UNet (Ronneberger, Fischer, and Brox 2015), and stacked HG (Newell, Yang, and Deng 2016). On the other hand, some focus on enhancing loss function. For example, Awing (Wang, Bo, and Fuxin 2019) addresses the sample imbalance problem between foreground and background in heatmaps. Beyond the point-based heatmap, several studies leverage edges to enhance the accuracy of contours (Wu et al. 2018; Huang et al. 2020, 2021).

Data Synthesis. For generating plenty of facial landmark data and annotations in model training, (Robinson et al. 2019; Qian et al. 2019) utilize GAN, and (Huang and Tamrakar 2020; Browatzki and Wallraven 2020) leverage 3D reconstruction. Both of them turned out to be effective in the lack of data. Whereas, most of them depend on auxiliaries, such as data in other domains and 3D models.

Dataset Construction. A number of face alignment datasets are published with diverse scenarios and various annotation definitions, but they rarely contain denser landmark annotations than 100, such as 21 landmarks in AFLW (Koestinger et al. 2011), 29 landmarks in COFW (Burgos-Artizzu, Perona, and Dollár 2013), 68 landmarks in 300W (Sagonas et al. 2013) and 98 landmarks in WFLW (Wu et al. 2018), due to costly annotation.

Method

We propose a weakly-supervised framework, called FreeEnricher, which can enrich the facial landmarks by given sparse landmarks. Motivated by our observation that the appearance of local patches on each facial contour are similar, the main idea is proposed that learning the refinement ability of original sparse landmarks on local patches and transfering this ability to enriched dense landmarks. Finally, the FreeEnricher can be plug-and-play to existing face alignment networks and enrich landmarks to arbitrary density without additional cost, e.g. denser landmark annotation, and more inference time.

FreeEnricher Framework

To construct the unified framework, several operators are devised and organized together in Fig. 2. They are *landmark initializing*, *offset generating*, *patch cropping*, *patch normalizing*, *quality scoring*, and *index embedding*. A detailed description of each operator could be referred to below.

Landmark Initializing initialize the rough enriched landmarks from sparse landmarks by averagely interpolating points from each facial contour, such as eye, mouth, and face contour. Hence, straight-line fitting and b-spline fitting (Knott 2000) could be selectively applied to fit the contours,

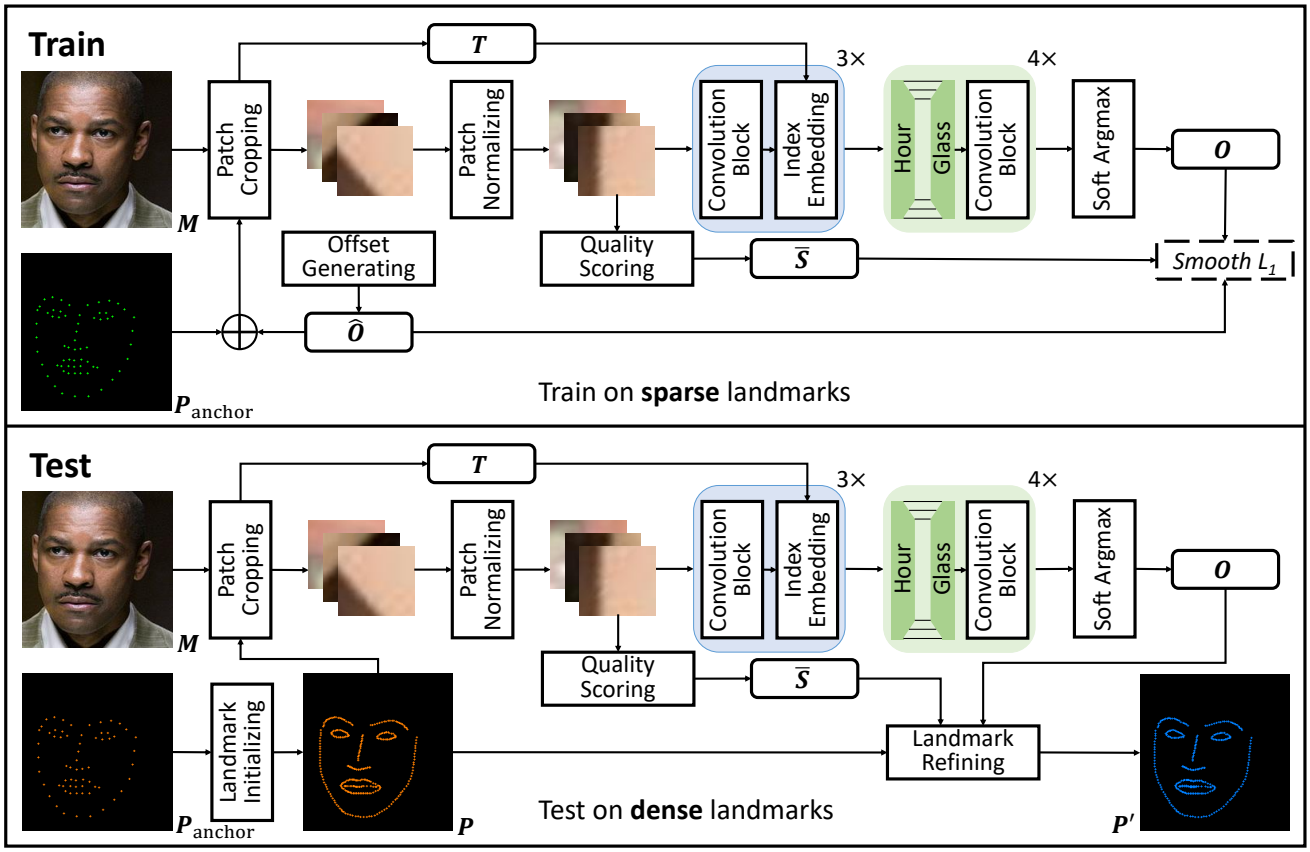


Figure 2: The framework of FreeEnricher. The framework employs the proposed weakly-supervised idea to enrich landmarks that training on the original sparse landmarks and testing on the enriched dense landmarks. In training stage, as shown in the upper part of the figure, a random offset is supervised on each normalized patch per sparse landmark and applies the normalized quality score as weight in loss function. In testing stage, as shown in the lower part of the figure, the enriched landmarks are initialized firstly, then refined by the trained model for each of them. Briefly explaining symbols, M is the face image, P_{anchor} is the original landmarks, P is the initialized enriched landmarks, P' is the final enriched landmarks, $\hat{O}$ is the randomly generated offset, O is the regressed offset, T is the soft index, and $\bar{S}$ is the normalized quality score.

and we call each fitted contour C. Given the enriching density D of *landmark initializing* and the original landmarks P_{anchor}, the rough enriched landmarks P of each contour are indicated as

$$p_{i,j} = \begin{cases} C(u_{i,j}), & j \in \{1, 2, ..., D-1\} \\ C(u_i), & j = 0 \end{cases} \quad (1)$$

where i is the index of *anchor point*, j is the index of *interpolated point* between two adjacent *anchor points* and u controls the relative position of enriched point in curve. We denote the original landmark as *anchor point*, where $j = 0$, and the newly sampled point as *interpolated point*, where $j > 0$. Then the u of *interpolated point* is derived by

$$u_{i,j} = \frac{D-j}{D} \cdot u_i + \frac{j}{D} \cdot u_{i+1} \quad (2)$$

where u_i and u_{i+1} present two adjacent *anchor points*. Since the curve is derived by only considering the existing sparse landmarks in geometry, the newly generated points are still not consistent with the image texture.

Offset Generating simulates the small position errors of *interpolated points* in normal by applying a random 2D offset for each *anchor point* in normal direction during training stage. We denote the generated offset by $\hat{O}$, which is the ground truth. Given the local patch size s_{patch}, the offset follows the uniform distribution $U(-s_{patch}/8, +s_{patch}/8)$.

Patch Cropping crops local image patch from the face image at the center of input point by the size of $s_{patch} \times s_{patch}$, where the input point denotes distorted anchor points $P_{anchor} + \hat{O}$ in training stage and initialized enriched landmarks P in testing stage. The relative scale of cropped patch is controlled by the patch-face ratio which represents the ratio of patch size to aligned face size.

Patch Normalizing rotates each patch so that its x-axis (horizontal direction) is aligned with the normal of the corresponding landmark along the contour line, which is illustrated in the left 2 columns of Fig. 3. The operator is based on the assumption that the precision improvement along the normal direction is more effective and reasonable than the tangent direction (Huang et al. 2021). Furthermore, the freedom of the regression target is reduced from 2D to 1D for no ambiguity and better convergence.

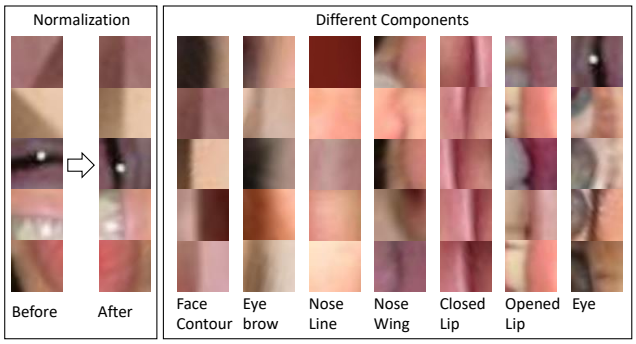


Figure 3: Patch samples that 64×64 region is cropped from 1024×1024 aligned face image (patch-face ratio is 1/16). The left 2 columns show the patches before/after normalization. The right 7 columns show the typical patches from different facial components.

Quality Scoring is designed to handle blurry boundaries, occluded regions, and contour corners. The goal of this operator is to provide the reliability of each patch, which is used to weight the losses in training and provide confidence in testing. The intermediate value V of the score is derived by

$$V(k) = \frac{\sigma(\sum_{y=1}^{h} patch_k(x,y))}{\sigma(\sum_{x=1}^{w} patch_k(x,y))} \tag{3}$$

where w and h are the width and height of the local image patch, $patch_k(x,y)$ is the gray value of the specific pixel in kth patch and $\sigma(\cdot)$ is the standard deviation. Equation 3 is maximized when the variance along the horizontal direction is high, and the variance along the vertical direction is low, which aims to find a clear vertical boundary in the aligned image patch, in other words, those patches are suitable to regress the accurate rectification of the landmark because of the similar appearance and clear edge feature. The raw score S of the patch is then obtained by

$$S(k) = \begin{cases} V(k) - 1, & V(k) >= 1 \\ 1 - \frac{1}{V(k)}, & otherwise \end{cases} \tag{4}$$

To normalize the raw score within 0 and 1, we firstly build a cumulative distribution function (CDF) using the whole patch scores in the training set, then map each raw score to the final normalized score $\bar{S}$ by the cumulated probability in Equation 5.

$$\bar{S}(k) = \frac{1}{N} \cdot \sum_{n=1}^{N} \begin{cases} 1, & S(n) < S(k) \\ 0, & otherwise \end{cases} \tag{5}$$

where N is the total number of patches in the training dataset. The distribution of the quality scores and their patch samples are shown in Fig. 4, which verifies the rationality in visual. In general, the aligned patch with a high-quality score has a clear vertical boundary, with a medium-quality score has a blurry boundary or occlusion, and with a low-quality score has a contour corner.

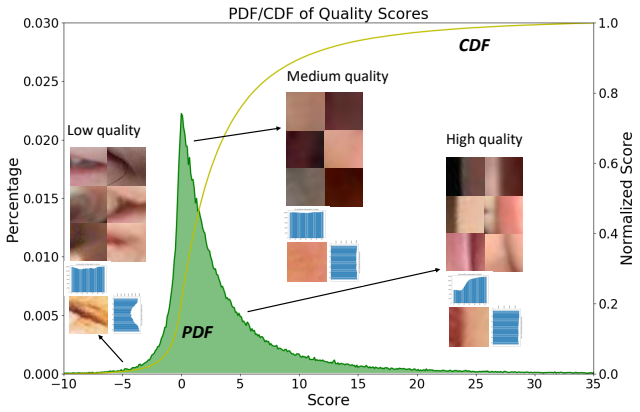


Figure 4: PDF/CDF of patch quality score in 300W training dataset. The green distribution stands for the probability density function (PDF) of scores whose y-axis is at left, and the yellow curve indicates the cumulative distribution function (CDF) of scores whose y-axis is at right. The CDF is also the mapping function from score to normalized score. The patches are divided into low-quality, medium-quality, and high-quality category. Their corresponding typical patch samples are exhibited by category. Meanwhile, the distributions of pixel values in column and in row are also presented for a better understanding of the quality scoring.

Index Embedding provides the position code of each patch similarity to transformer network (Vaswani et al. 2017). Specifically, we assign a soft index (position) to each patch (word) on the face (sentence).The right 7 columns of Fig. 3 show that the patches in the same component are similar, but there still exists a marginal appearance gap across components. Meanwhile, we can not put the patches of each landmark to the model in order, because of the misaligned landmark density between training and testing. For the above reasons, the *index embedding* in our framework is necessary to further improve the generalization of the model.

Assuming that the feature maps $\mathbf{F}_{in}$ from the previous CNN layer have m channels, then $\{\lfloor t \rfloor, n + \lfloor t \rfloor, 2n + \lfloor t \rfloor, ..., m - n + \lfloor t \rfloor\}$ indices are selected as $\mathbf{U}$ to filter channels from input feature maps, where n is the number of *anchor points*. Given the soft index t, the output feature maps $\mathbf{F}_{out}$ of *index embedding* are presented as

$$\mathbf{F}_{out} = (1 + \lfloor t \rfloor - t) \cdot \mathbf{F}_{in}(\mathbf{U}) + (t - \lfloor t \rfloor) \cdot \mathbf{F}_{in}(\mathbf{U} + 1) \tag{6}$$

To bridge the different landmark density between the training and testing stages of FreeEnricher, the soft index t of $p_{i,j}$ could be calculated by

$$T(i,j) = i + \frac{j}{D} \tag{7}$$

where $p_{i,j}$ is the enriched landmark. Specifically, $p_{i,j}$ is the anchor point, when $j = 0$, otherwise, the interpolated point.

Landmark Refining refines the initialized enriched landmarks P by adding the regressed offset O with normalized quality score $\bar{S}$ as confidence (weight), which denotes as

$$P' = P + \bar{S} \times O \tag{8}$$

where P' is the final refined enriched landmarks by FreeEnricher in inference.

Training process of FreeEnricher is presented in the upper part of Fig. 2, by inputting the face image M and original landmarks P_{anchor}, we firstly apply the *offset generating* to the original landmarks P_{anchor} to generate the ground truth of offsets $\hat{O}$, which would distort the original landmarks P_{anchor}, then feed the face image M with distorted original landmarks $P_{anchor} + \hat{O}$ to model (start from *patch cropping* and end in *soft argmax*) to get the regressed offsets O and normalized quality scores $\bar{S}$, finally put them ($\bar{S}$, O, $\hat{O}$) to the loss function of FreeEnricher $\mathcal{L}$, which is defined as

$$\mathcal{L} = \frac{1}{B} \cdot \sum_{i=1}^{B} \bar{S}(i) \cdot Smooth\ L_1(O(i), \hat{O}(i)) \qquad (9)$$

where B is the batch size, O is the regressed offset, $\hat{O}$ is the generated offset, and $\bar{S}$ is the normalized quality score.

Testing process of FreeEnricher is demonstrated in the lower part of Fig. 2, by inputting the face image M and original landmarks P_{anchor}, we firstly employ the *landmark initializing* on the original landmarks P_{anchor} to generate initialized enriched landmarks P, then feed the face image M with initialized enriched landmarks P to model (start from *patch cropping* and end in *soft argmax*) to get the regressed offsets O and normalized quality scores $\bar{S}$, finally put them (P, $\bar{S}$, O) to *landmark refining* to acquire the final accurate enriched landmarks P'.

FreeEnricher Network

The model from FreeEnricher framework has the ability to enrich sparse landmarks to accurate dense landmarks. To make full use of this ability, we apply the FreeEnricher model to the existing face alignment networks (baseline networks) by plug-and-play mode, then get the FreeEnricher networks (enhanced networks). Specifically, there are two ways to plug in the model: (1) employing the model to preprocess the training dataset, which means densifying the ground truth of sparse landmarks, and (2) utilizing the model to postprocess the testing results, which denotes densifying the predicted value of sparse landmarks. Each way can be used optionally, thus we define the following 3 modes.

Plug in Test denotes that FreeEnricher only enriches the predicted landmarks in the testing stage. Under this mode, the total inference time increases tremendously.

Plug in Train indicates that FreeEnricher merely enriches the ground truth of landmarks in the training stage. Under this mode, there is no extra computational cost in inference, and it is the default mode in our method.

Plug in Train and Test means that FreeEnricher model enriches the ground truth of landmarks in the training stage with *landmark initializing* and refines the predicted landmarks in the testing stage without *landmark initializing*. Under this mode, the network always gets the highest accuracy, but it is pretty time-consuming in practical deployment.

FreeEnricher Network	Baseline Network	Enriching Density	Plug in Train	Plug in Test
ResNet50-FE2$_{train}$	ResNet50	2	✓	✗
HRNet-FE3$_{test}$	HRNet	3	✗	✓
ADNet-FE5	ADNet	5	✓	✗
ADNet-FE10$_{train+test}$	ADNet	10	✓	✓

Table 1: FreeEnricher network samples.

The name of FreeEnricher network is defined as "$BaselineNet$-FED_{stage}" in a unified format, where $BaselineNet$ indicates which face alignment baseline network to be used, D denotes the enriching density of *landmark initializing* and $stage$ means in which stage the FreeEnricher plug, $stage$ could be omitted when its value is "train". Several FreeEnricher network samples are presented in Table 1 for clearer understanding.

Experiment

Implementation Details

FreeEnricher. The FreeEnricher model is trained on four GPUs (16GB NVIDIA Tesla P100) by PyTorch (Paszke et al. 2019), where the batch size of each GPU is 68 for 300W and 98 for WFLW. We employ Adam optimizer with the initial learning rate of 1×10^{-3} and decay the learning rate by $1/10$ at the epochs of 100, 150, and 180, finally ending at 200. Specifically, *landmark initializing* adopts b-spline interpolation method implemented in Scipy (Virtanen et al. 2020) with order of 3 and enriching density of 5, *offset generating* randomly generates offset by the uniform distribution $U(-8, +8)$, and *patch cropping* crops 64×64 patch from aligned face image with 1024×1024 resolution (patch-face ratio of 1/16) by the center of target landmark. The normalized patches are augmented by simulated random gray, random blur, and random occlusion. Regarding the backbone architecture of FreeEnricher, we adopt 3-stacked *index embedding* modules and 4-stacked *hourglass* modules, where each *hourglass* outputs a $1 \times 64 \times 64$ feature map. The output feature maps are then fed to $soft\ argmax$ (Huang et al. 2021) to generate the landmark offsets. The final landmark offset is from the last *hourglass*. The following experiments employ the settings above if no further explanation.

Face Alignment Network. The FreeEnricher model enhances the existing face alignment network by "Plug in Train" mode in both landmark density and accuracy. And all of the existing face alignment networks are trained using the identical setting to the original papers.

Enriched 300W test set. We newly construct an enriched 300W test dataset by manually labeling the original images, which is employed to evaluate the performance of enriched face alignment. To handle different densities of landmarks, we labeled the continuous curve rather than discrete points. From the curve, firstly, we extract the anchor points by finding the points (on the curve) which are closest to the original landmark labels in 300W. Then, between two neighboring anchor points, the new points are uniformly sampled.

Figure 5: Annotation samples of the enriched 300W benchmark with 320 landmarks. Those annotations are newly labeled by experts.

For instance, from the original 68 landmarks, 320 enriched landmarks are generated with the enriching density 5. Typical samples are shown in Fig. 5.

Evaluation Metrics

Mean Error (ME) measures the averaged $L2$ distance between predicted points and ground-truth points, which is specifically employed to evaluate the offset regression of fixed-scale patch in FreeEnricher model and defined as

$$\mathrm{ME}(P, \hat{P}) = \frac{1}{N_P} \sum_{i=1}^{N_P} \|p_i - \hat{p}_i\|_2 \tag{10}$$

where P and $\hat{P}$, respectively, denotes the predicted values and the ground-truth of landmarks (or offsets), and N_P is the number of landmarks (or offsets).

Normalized Mean Error of Points (NME_{point} or NME) between predicted points and ground-truth points is a widely-used standard metric to evaluate landmark accuracy for face alignment, which is defined as

$$\mathrm{NME}_{point}(P, \hat{P}) = \frac{\mathrm{ME}(P, \hat{P})}{d} \tag{11}$$

where d is the unit distance used to normalize the errors. Inter-ocular distance (distance between outer eye corners) is employed as the unit distance in the following experiments.

Normalized Mean Error of Edges (NME_{edge}) between predicted points and ground-truth edges is proposed to evaluate the performance of the landmarks, which emphasizes the error on the normal direction and is defined as

$$\begin{aligned}\mathrm{NME}_{edge}(P, \hat{P}, \hat{E}) = &\frac{1}{N_P} \sum_{i=1}^{\hat{p_i} \notin \hat{E} \ and \ N_P} \frac{\|p_i - \hat{p}_i\|_2}{d} + \\ &\frac{1}{N_P} \sum_{i=1}^{\hat{p_i} \in \hat{E} \ and \ N_P} \frac{dist(p_i, \hat{e}_i)}{d}\end{aligned} \tag{12}$$

where $\hat{E}$ is the ground-truth of edges, $\hat{e}_i$ is the edge that the ith landmark belongs to, and $dist(\cdot)$ computes point-to-edge distance.

Comparison on Enriched Landmarks

The main advantage of applying the FreeEnricher is that we get denser landmarks than from the original dataset. In order to demonstrate the effectiveness of FreeEnricher on enriching landmarks, we evaluate the NME of landmarks in the newly-constructed enriched 300W testing dataset. Because there is no existing method that could freely enrich landmarks, we design a baseline method as the comparison by applying *landmark initializing* operator, which enriches the landmark density of the original network. Meanwhile, the state-of-the-art ADNet is selected as the baseline network to be enhanced for a good starting point. Besides, other settings are identical between these 2 experiments. The result could be referred to Table 2, where FreeEnricher significantly enhances the accuracy of enriched landmarks, especially $\sim$20% improvement in the normal direction.

Method	Network	NME_{point}	NME_{edge}
Baseline	ADNet + Line5	3.21	1.18
Ours	**ADNet-FE5**	**3.06**	**0.98**

Table 2: Comparison between FreeEnricher and baseline method on the enriched 300W testing set, totally 320 landmarks. "Line5" represents applying *landmark initializing* to the output of network by straight-line interpolation with D of 5.

Comparison on Original Landmarks

Beyond enriching landmarks, the FreeEnricher can benefit the baseline network on the sparse landmarks as well. For the same reason, we still use ADNet as the baseline. Hence, we also evaluate FreeEnricher and the other face alignment algorithms on the original 300W and WFLW datasets in Table 3. The results show that FreeEnricher unexpectedly improves the accuracy further, and outperforms the existing methods on both 300W and WFLW.

Method	300W	WFLW
LAB (Wu et al. 2018)	3.49	5.27
HRNet (Wang et al. 2020)	3.34	4.60
LUVLi (Kumar et al. 2020)	3.23	4.37
ADNet (Huang et al. 2021)	2.93	4.14
ADNet-FE5 (Ours)	**2.87**	**4.10**

Table 3: NME comparison between our method and other state-of-the-art on the original 300W and WFLW testing set, 68 and 98 landmarks in total respectively. To compare the result of ADNet-FE5 network, we merely select the original landmarks from the enriched landmarks.

Ablation Study

To verify the efficacy of FreeEnricher in unit and the contribution of each operator, we conduct comprehensive ablation experiments. Furthermore, they also demonstrate the robustness of different face alignment networks and the applicability to various landmark densities.

Unit Verification. To evaluate the offset refinement network independently from the whole system on local image patches, we conduct unit verification in Table 4. Hereon, FreeEnricher is trained on the generated patches from the training set and tested from the testing set respectively. Both the results on 300W and WFLW datasets demonstrate that FreeEnricher could refine the landmark location in the local patch of the face image.

Method	300W Testset	WFLW Testset
Random offset	4.03	3.96
Regressed offset	**1.92**	**1.89**

Table 4: Unit verification of FreeEnricher on 300W and WFLW datasets. The second row indicates the ME of the generated offsets by *offset generating* operator. The third row indicates the ME of the regressed offsets by FreeEnricher.

Contribution of Each Operator. After integrating the designed operators into the framework individually, the performance increases to different degrees, which demonstrates the contribution of each operator in FreeEnricher framework. B-spline *landmark initializing* and *patch normalizing* improve more significantly than *quality scoring* and *index embedding* on accuracy. The first setting in Table 5 indicates average upsampling original landmarks by linear interpolation without FreeEnricher regression, which is the baseline. And the last setting in Table 5 achieved the best result.

Landmark Initializing	Patch Normalizing	Quality Scoring	Index Embedding	NME
-	-	-	-	3.21
line	✗	✗	✗	3.18
b-spline	✗	✗	✗	3.11
line	✓	✗	✗	3.14
line	✗	✗	✓	3.16
line	✓	✓	✗	3.11
b-spline	✓	✓	✓	**3.06**

Table 5: Operator contribution on enriched 300W dataset.

Robustness on Existing Network. The FreeEnricher could be plug-and-play to any face alignment network in theory. To verify the robustness in practice, several networks, ResNet50 (Zhang et al. 2014b), HRNet (Wang et al. 2020), and ADNet (Huang et al. 2021), are selected to conduct the comparison experiments. Table 6 demonstrate that all of them have a significant improvement on enriched 300W testing set after applying the FreeEnricher.

Applicability on Enriching Density The FreeEnricher could enrich landmarks to arbitrary density in theory. To verify the applicability in practice, various enriching densities, 1, 2, 3, 5, and 10, are selected to conduct the verification experiments. Table 7 demonstrates that all of them achieve consistently high accuracy on the enriched 300W testing dataset by the given enriching density. Besides, the denser the landmarks are, the larger the NME becomes.

Network	NME
ResNet50 + Line5 (Zhang et al. 2014b)	4.76
ResNet50-FE5 (Ours)	**4.39**
HRNet + Line5 (Wang et al. 2020)	3.74
HRNet-FE5 (Ours)	**3.46**
ADNet + Line5 (Huang et al. 2021)	3.21
ADNet-FE5 (Ours)	**3.06**

Table 6: Robustness study of FreeEnricher on different face alignment network on enriched 300W testing set.

Enriching Density	Landmarks Count	NME
1	68	2.91
2	131	3.01
3	194	3.03
5	320	3.06
10	635	3.14

Table 7: Applicability study of FreeEnricher on different enriching density D on enriched 300W testing dataset.

Study on Plugging Stage. The FreeEnricher could be plug-and-play in existing networks under their training and testing stage. The experiments in Table 8 illustrate that the FreeEnricher plugged in different stages have similar accuracy improvement but different computational cost in inference. To balance both high accuracy and low computational cost in the deployment environment, FreeEnricher is merely plugged in training stage. Then, FreeEnricher brings improvement on enriched landmarks without extra inference time and memory.

Network	NME	Extra FLOPs	Extra Params
ADNet-FE5$_{test}$	3.07	+9.85 G	+6.78 M
ADNet-FE5$_{train}$	3.06	**+0 G**	**+0 M**
ADNet-FE5$_{train+test}$	**3.04**	+9.85 G	+6.78 M

Table 8: Study of FreeEnricher on different plugging stage on enriched 300W testing dataset. The subscript of the network name denotes the plugging stage of FreeEnricher.

Conclusion

Enriching facial landmarks is a meaningful topic in both academic and industry. In this paper, we present a novel idea to address this problem by using the existing datasets without additional labeling annotation and inference time costs. Moreover, our method could enrich landmarks to arbitrary density and be plug-and-play to any existing face alignment network in theory, which is also demonstrated by various experiments. Specifically, a weakly-supervised framework is proposed, which learns the refinement ability on original sparse landmarks of existing datasets, then this ability is applied to the initialized dense landmarks. By our method, we not only obtain accurate enriched facial landmarks but also achieve state-of-the-art performance in the original dataset.

References

Browatzki, B.; and Wallraven, C. 2020. 3fabrec: Fast few-shot face alignment by reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6110–6120.

Burgos-Artizzu, X. P.; Perona, P.; and Dollár, P. 2013. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE international conference on computer vision*, 1513–1520.

Cootes, T. F.; Edwards, G. J.; and Taylor, C. J. 2001. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6): 681–685.

Cootes, T. F.; and Taylor, C. J. 1992. Active shape models—'smart snakes'. In *BMVC92*, 266–275. Springer.

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

Deng, J.; Trigeorgis, G.; Zhou, Y.; and Zafeiriou, S. 2019. Joint multi-view face alignment in the wild. *IEEE Transactions on Image Processing*, 28(7): 3636–3648.

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

Feng, Z.-H.; Kittler, J.; Awais, M.; Huber, P.; and Wu, X.-J. 2018. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2235–2245.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Gundavarapu, N. B.; Srivastava, D.; Mitra, R.; Sharma, A.; and Jain, A. 2019. Structured Aleatoric Uncertainty in Human Pose Estimation. In *CVPR Workshops*, volume 2, 2.

Huang, J.; and Tamrakar, A. 2020. ACE-Net: Fine-Level Face Alignment through Anchors and Contours Estimation. *arXiv preprint arXiv:2012.01461*.

Huang, X.; Deng, W.; Shen, H.; Zhang, X.; and Ye, J. 2020. PropagationNet: Propagate Points to Curve to Learn Structure Information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7265–7274.

Huang, Y.; Yang, H.; Li, C.; Kim, J.; and Wei, F. 2021. ADNet: Leveraging Error-Bias Towards Normal Direction in Face Alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3080–3090.

Jiang, D.; Hu, Y.; Yan, S.; Zhang, L.; Zhang, H.; and Gao, W. 2005. Efficient 3D reconstruction for face recognition. *Pattern Recognition*, 38(6): 787–798.

Knott, G. D. 2000. *Interpolating cubic splines*, volume 18. Springer Science & Business Media.

Koestinger, M.; Wohlhart, P.; Roth, P. M.; and Bischof, H. 2011. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, 2144–2151. IEEE.

Kumar, A.; Marks, T. K.; Mou, W.; Wang, Y.; Jones, M.; Cherian, A.; Koike-Akino, T.; Liu, X.; and Feng, C. 2020. LUVLi Face Alignment: Estimating Landmarks' Location, Uncertainty, and Visibility Likelihood. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8236–8246.

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

Liu, Y.; Jourabloo, A.; Ren, W.; and Liu, X. 2017. Dense face alignment. In *IEEE International Conference on Computer Vision Workshops*, 1619–1628.

Luo, P.; Wang, X.; and Tang, X. 2012. Hierarchical face parsing via deep learning. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2480–2487. IEEE.

Lv, J.; Shao, X.; Xing, J.; Cheng, C.; and Zhou, X. 2017. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3317–3326.

Masi, I.; Wu, Y.; Hassner, T.; and Natarajan, P. 2018. Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, 471–478. IEEE.

Matthews, I.; and Baker, S. 2004. Active appearance models revisited. *International journal of computer vision*, 60(2): 135–164.

Milborrow, S.; and Nicolls, F. 2008. Locating facial features with an extended active shape model. In *European conference on computer vision*, 504–513. Springer.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, 483–499. Springer.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Qian, S.; Sun, K.; Wu, W.; Qian, C.; and Jia, J. 2019. Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10153–10163.

Robinson, J. P.; Li, Y.; Zhang, N.; Fu, Y.; and Tulyakov, S. 2019. Laplace landmark localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10103–10112.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; and Pantic, M. 2013. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 397–403.

Saragih, J.; and Goecke, R. 2007. A nonlinear discriminative approach to AAM fitting. In *2007 IEEE 11th International Conference on Computer Vision*, 1–8. IEEE.

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

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

Trigeorgis, G.; Snape, P.; Nicolaou, M. A.; Antonakos, E.; and Zafeiriou, S. 2016. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4177–4187.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, 17(3): 261–272.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*.

Wang, X.; Bo, L.; and Fuxin, L. 2019. Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6971–6981.

Wei, S.-E.; Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2016. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 4724–4732.

Wu, W.; Qian, C.; Yang, S.; Wang, Q.; Cai, Y.; and Zhou, Q. 2018. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2129–2138.

Zhang, J.; Shan, S.; Kan, M.; and Chen, X. 2014a. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European conference on computer vision*, 1–16. Springer.

Zhang, Z.; Luo, P.; Loy, C. C.; and Tang, X. 2014b. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, 94–108. Springer.

Zhou, E.; Fan, H.; Cao, Z.; Jiang, Y.; and Yin, Q. 2013. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE international conference on computer vision workshops*, 386–391.

Zhu, X.; Lei, Z.; Liu, X.; Shi, H.; and Li, S. Z. 2016. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 146–155.