applied sciences

MDPI

Article

Learning Robust Shape-Indexed Features for Facial Landmark Detection

Xintong Wan, Yifan Wu and Xiaoqiang Li *

School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China; wanxintong@shu.edu.cn (X.W.); victorwu@shu.edu.cn (Y.W.)
* Correspondence: xqli@shu.edu.cn

Abstract: In facial landmark detection, extracting shape-indexed features is widely applied in existing methods to impose shape constraint over landmarks. Commonly, these methods crop shape-indexed patches surrounding landmarks of a given initial shape. All landmarks are then detected jointly based on these patches, with shape constraint naturally embedded in the regressor. However, there are still two remaining challenges that cause the degradation of these methods. First, the initial shape may seriously deviate from the ground truth when presented with a large pose, resulting in considerable noise in the shape-indexed features. Second, extracting local patch features is vulnerable to occlusions due to missing facial context information under severe occlusion. To address the issues above, this paper proposes a facial landmark detection algorithm named Sparse-To-Dense Network (STDN). First, STDN employs a lightweight network to detect sparse facial landmarks and forms a reinitialized shape, which can efficiently improve the quality of cropped patches when presented with large poses. Then, a group-relational module is used to exploit the inherent geometric relations of the face, which further enhances the shape constraint against occlusion. Our method achieves 4.64% mean error with 1.97% failure rate on COFW68 dataset, 3.48% mean error with 0.43% failure rate on 300 W dataset and 7.12% mean error with 11.61% failure rate on Masked 300 W dataset. The results demonstrate that STDN achieves outstanding performance in comparison to state-of-the-art methods, especially on occlusion datasets.

Keywords: facial landmark detection; shape-indexed feature; face shape constraint; biometrics

Citation: Wan, X.; Wu, Y.; Li, X. Learning Robust Shape-Indexed Features for Facial Landmark Detection. *Appl. Sci.* **2022**, *12*, 5828. https://doi.org/10.3390/app12125828

Academic Editors: Monica Perusquia Hernandez and Saho AYABE-Kanamura

Received: 29 April 2022
Accepted: 5 June 2022
Published: 8 June 2022

1. Introduction

Facial landmark detection, also known as face alignment, aims to localize landmarks of given faces. It is an essential step in many face analysis tasks, e.g., face verification [1–3], expression recognition [4–6], face editing [7,8] and face recognition [9,10].

In recent years, convolutional neural networks (CNNs) have promoted the progress of robust facial landmark detection. However, the robustness of landmark detection on unconstrained faces still suffers from occlusion, illumination and large pose variation problems.

To achieve robust facial landmark detection, some works [11–13] impose face shape constraint over all landmarks against occlusion. For example, LAB [11] imposes the shape constraint by estimating the boundary information that is predicted by an additional stacked hourglass network. However, facial boundary estimation significantly increases computational costs. Other methods, such as MDM [12], learn the shape-indexed features from local patches surrounding a mean shape to predict all landmarks, and the shape constraint is encoded in the regressor. Figure 1 shows the local patches used to learn shape-indexed features in existing methods. Figure 1a,b shows the problems with two initialization strategies when presented with a large pose. The initial landmarks are extremely far from the ground-truth landmarks. In addition, shape-indexed features only provide coarse shape constraints, which are vulnerable to occlusion due to the lack of facial context in local patches.

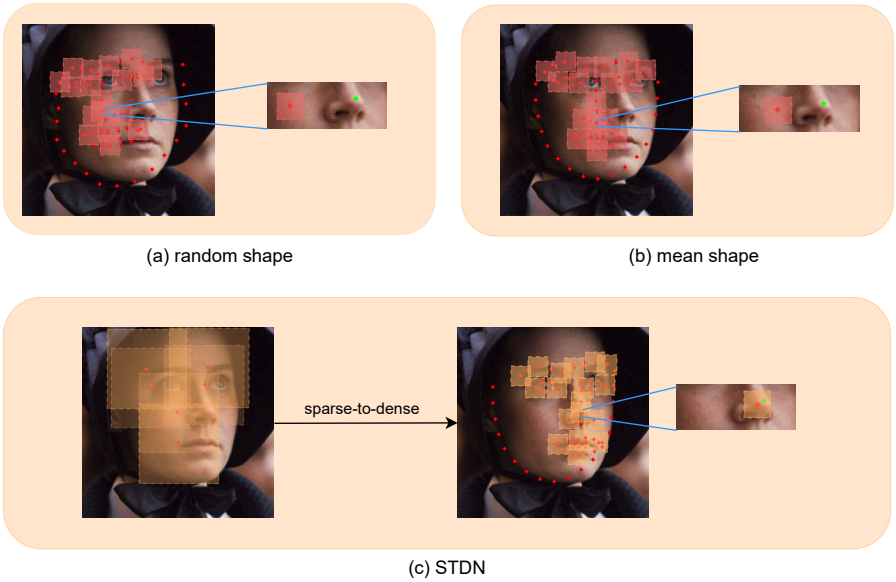

Figure 1. Shape-indexed local patches of three sampling approaches where red points represent the initial landmarks and green points represent the ground-truth landmarks. (**a**,**b**) show the sampling result using the random shape and mean shape, respectively, as initial landmarks on a face with large pose variation. (**c**) shows our sampling approach, in which the local patches capture the ground-truth landmarks more precisely than the other two approaches.

This paper proposes a sparse-to-dense network (STDN) to reduce the noise data under large pose variations and handle the occlusion problem in facial landmark detection. The process is functionally divided into two stages: the patch resampling stage and the relation reasoning stage. In the patch resampling stage, STDN adopts the sampling method as shown in Figure 1c. First, STDN downsamples the mean shape into sparse landmarks and then crops large-sized local patches by using these sparse landmarks. This allows us to use a lightweight network to predict a set of offset values based on these large-sized patches. Then, according to these offsets, the mean shape is adjusted to a reinitialized shape. In the relation reasoning stage, the input is the small-sized local patches cropped surrounding the reinitialized shape. The whole features, learned based on such small-sized patches, are used to predict the whole face shape. A group-relational module exploits the geometric relations between facial components, which first disentangles the nose feature from all features to constrain the other facial components according to the geometric relations. Meanwhile, all features play a role in imposing the global shape constraint. The main contributions of this work are summarized as follows:

- We propose a sparse-to-dense network (STDN), a two-stage framework, to reduce the noise data with large pose variations and address the severe occlusion problem;
- We suggest a sparse to dense patch sampling strategy to efficiently improve the quality of the cropped local patches with large pose variations;
- We take advantage of a group-relational module to handle the severe occlusion problem, which learns the geometric relations between facial components to enhance the shape constraint against occlusion.

2. Related Work

Facial landmark detection falls into three main categories, i.e., classic methods, coordinate regression methods and heatmap regression methods. Although these methods have achieved great success, it is still challenging to deal with severe occlusion and large pose variations.

Classic methods, such as ASM [14] and AAM [15], are based on statistical shape models. They use the principal component analysis (PCA) method to model the appearance and

shape by updating the coefficient vector, which can minimize the difference between shape-based appearance and input images. However, these methods only rely on the appearance features so that the performances of models tend to severely degrade when dealing with occlusion and faces with large pose variations.

Coordinate regression methods directly predict the coordinates of landmarks from the input image using regression models without relying on appearance models. These methods [12,16–22] typically utilize a coarse-to-fine manner to update the shape iteratively. DR [18] used a global layer to estimate the initial shape and then uses multiple local layers to update the shape iteratively. Park et al. [20] pretrained a feature extraction network to learn local feature descriptors from global facial features, which led to a higher face alignment accuracy. TR-DRN [21] designed a two-stage network to solve the initialization issue, which used the full face region for rough prediction in the global stage and refined landmarks in different parts of the face in the local stage. DAC-CSR [22] separated the face into multiple domains to train the domain-specific cascaded shape regression (CSR). Then, it used the dynamic attention-controlled method to select the appropriate subdomain CSR for landmark refinement. Coordinate regression methods take a small amount of time and ameliorate the robustness of the classic methods when facing the easy occlusion, but are not robust enough to handle the severe occlusion.

Some regression methods [23–25] also learn regression models based on the shape-indexed features that were first proposed in ESR [23]. It used the mean shape as the initial shape and gradually updated the landmarks by predicting the offset based on the local features extracted surrounding the initial shape. Wu et al. [24] considered that different face shapes should have various regression functions. Therefore, the model they proposed can automatically change the regression parameters according to current face shapes to better approximate the ground-truth shapes.

Heatmap regression methods [11,26–32] obtain the heatmap by generating a Gaussian distribution over the channels; the point with the highest response on the predicted heatmap is liable to be the prediction. DU-Net [27] used a quantized densely connected U-Net for effective facial landmark localization and used a K-order dense connection to achieve better detection accuracy with fewer parameters. AWing [29] designed a loss function of heatmap regression that achieved a greater penalty for foreground pixels and a smaller penalty for background pixels. ADC [31] combined global and local feature information for facial landmark detection without sacrificing image resolution and quality. Heatmap regression methods can achieve good performance, but they require deep networks and many parameters, resulting in complicated calculations and slow detection.

In recent years, with more attention to severe occlusion and large pose variations, an increasing number of works [16,33–41] have aimed at overcoming such obstacles in facial landmark detection. RCPR [16] detected the occlusion area while estimating the landmarks, and used the occlusion proportion of the area to weight the regressor. PCD-CNN [33] took the detected 3D face pose as the initial condition to detect landmarks under large pose variations. ODN [34] achieved robustness for occlusion by applying adaptive weights to facial regions and restored low-rank features of occluded regions by exploiting the geometric structure of the face. LUVLI [35] used a stacked hourglass network to jointly estimate landmark locations, the uncertainties of these predicted locations, and the visibility of landmarks. CCDN [36] proposed a cross-order cross-semantic deep network to activate multiple related facial parts, which fully explored more discriminative and fined semantic features to solve the problems of partial occlusions and large pose variations. MTAAE [37] proposed a multi-task adversarial autoencoder network based on the idea of multi-task learning, which could learn the more representative facial appearance and improve face alignment performance in the wild. SAAT [38] proposed a sample-adaptive adversarial training approach, in which the attacker generated adversarial perturbations to reflect the weakness of the detector, and the detector must improve its robustness to adversarial perturbations to defend against adversarial attacks. DSCN [39] proposed a dual-attentional spatial-aware capsule network to improve the ability to capture the spatial positional

relations between landmarks by using the capsule network that can remember the location information of the entity. MSM [40] used spatial transformer networks, hourglass networks and exemplar-based shape constraint to detect landmark under unconstrained conditions. Fard et al. [41] designed two teacher networks, a Tolerant-Teacher and a Tough-Teacher, to guide the lightweight student network. The Tolerant-Teacher was trained using soft-landmarks created by active shape models, while the Tough-Teacher was trained using the ground truth landmarks. Meanwhile, they designed an assistive loss to determine the landmarks of teacher network prediction as positive or negative auxiliary.

3. Methods

As illustrated in Figure 2, the sparse-to-dense network mainly consists of two stages: the patch resampling stage and the relation reasoning stage. The first row of Figure 2 shows the patch resampling stage. This stage aims to improve the quality of shape-indexed patches cropped under large pose variations, which is beneficial for learning robust shape-indexed features. The second row of Figure 2 shows the relation reasoning stage, which exploits facial components' geometric relations to enhance the shape constraint in order to achieve robust detection on severe occlusion faces.

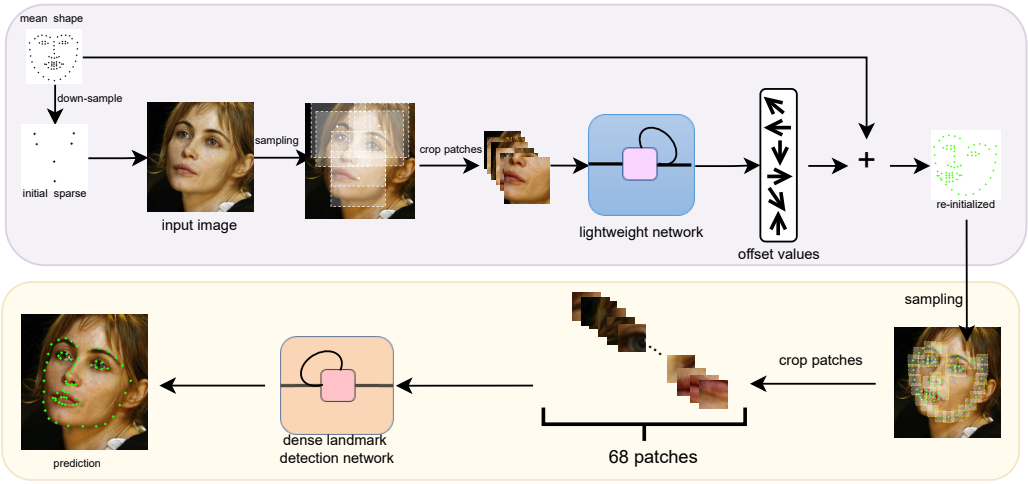

Figure 2. Overall architecture of the proposed network. It consists of two stages: the patch resampling stage (**the first row**) and the relation reasoning stage (**the second row**). The downsample operation in the first row averages the coordinates of the landmarks in each facial part according to the predefined indexes. The + operation in the first row adds the offset values of the corresponding sparse landmarks to the coordinates of all landmarks in each facial part of the mean shape.

3.1. Patch Resampling Stage

As shown in the first row of Figure 2, according to the predefined indexes, except for the cheek, the rest of the mean shape is downsampled into six landmarks corresponding to six facial parts: left eyebrow, right eyebrow, left eye, right eye, nose and mouth. This operation does not include the cheek because the landmarks of the cheek are distributed along the entire edge of the face and cannot be represented by a single point. These sparse landmarks allow us to crop large local patches that are fed into a lightweight network to acquire the offset values. Subsequently, all offset values are applied to the original mean shape to form a reinitialized shape, which is used to crop small shape-indexed patches. This resampling operation allows the STDN to improve the quality of shape-indexed patches, and thus, STDN extracts more robust shape-indexed features compared with existing methods.

Figure 3 depicts the sparse landmark detection diagram. The mean face shape is downsampled into six sparse landmarks as the initial shape S_0, and each landmark in this shape represents a facial part. Actually, it is the average of all landmarks in the specific

facial part. The offset values ΔS_n are predicted based on the shape-indexed patches of 100×100 size that are cropped surrounding the initial landmark. It is expressed as:

$$\Delta S_n = f(I, S_{n-1}), \tag{1}$$

where I is the input face image of size 384×456, S_{n-1} is the output shape of the last iteration, and $f(\cdot)$ represents the regression function of the lightweight network. The current shape S_n is obtained by updating S_{n-1} and used as the initial shape for the next iteration:

$$S_n = S_{n-1} + \Delta S_n. \tag{2}$$

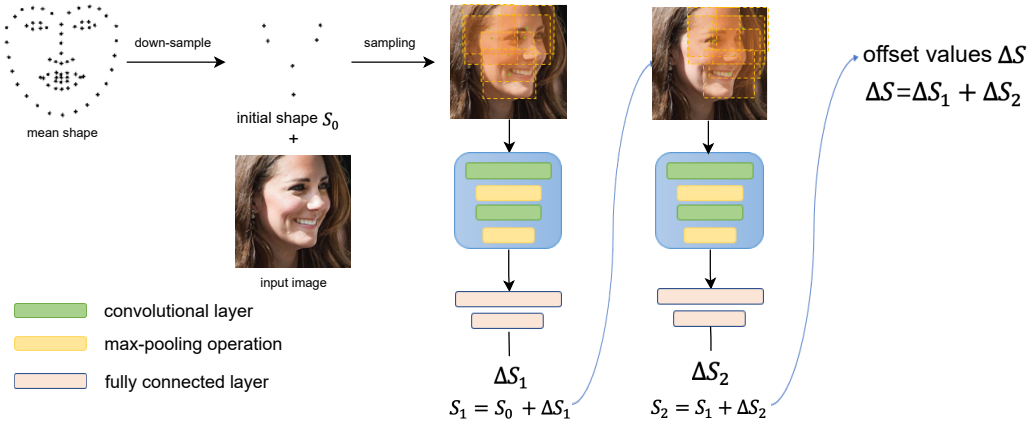

Figure 3. Sparselandmark detection diagram.

Take the sum of the offset of each iteration as the final offset values:

$$\Delta S = \sum_{n=1}^{N} \Delta S_n, \tag{3}$$

where N denotes the maximum number of iterations and is set to 2 in our implementation.

The parameters of this network are updated by minimizing the following loss function:

$$Loss = \sum_{n=1}^{N} \|\Delta S_n - (S^* - S_{n-1})\|_2^2, \tag{4}$$

where S^* represents the ground truth of sparse landmarks generated after the downsampling operation.

The reinitialized shape $\bar{S}$ is derived by applying the offset value of the corresponding sparse landmark to all landmarks in the facial part of the mean shape:

$$\bar{S} = \bigcup_{j=1}^{J} (S_0^j + \Delta S^j), \tag{5}$$

S_0^j is the j-th facial part of the mean shape, and $\bigcup$ denotes the operation of concatenating all six facial parts (i.e., both eyebrows, both eyes, nose and mouth). The landmarks of the cheek are taken directly from the mean shape. The reinitialized shape, which is used as the initial shape of the second stage is more similar to the ground truth than the mean shape.

3.2. Relation Reasoning Stage

An observation of unconstrained faces shows that the eyes and the eyebrows are often occluded by hair or sunglasses, the mouth may be occluded by food or microphone, and the cheek may be self-occluded due to large pose variations. Only the nose area is rarely

completely occluded, as it is located in the central region of the face. Considering all the above, we argue that the nose can be used as an anchor to constrain other facial components. The facial components are divided into six groups based on the inherent structure: the left group, including the left eyebrow and the left eye; the right group, including the right eyebrow and the right eye; the nose group; the mouth group the left cheek group; and the right cheek group, as shown in Figure 4.

Figure 4. Diagram of face division. Landmarks of the same color are divided into a group.

The relation reasoning stage employs a dense landmark detection network to exploit the shape constraint against occlusion, as shown in Figure 5. In this stage, the dense landmark detection network predicts the offset value at each iteration so that the initial shape will be gradually updated through multiple iterations. At the first iteration, the dense landmark detection network uses two candidate shapes to initialize: the reinitialized shape or the mean shape. The mean shape is used to avoid overfitting caused by heavily relying on the reinitialized shape. The probability of the reinitialized shape and the mean shape being chosen is ϵ and $1-\epsilon$, respectively. At the subsequent iterations, the prediction from the previous iteration is used for initialization. After obtaining the dense initial landmarks, we feed the cropped small-sized patches into the dense landmark detection network. This network uses three convolutional layers to extract shape-indexed features from small-sized patches. A fully connected layer is employed to transform shape-indexed features into whole features. The whole features are used to directly predict all landmarks; they are also used to combine with the shape-indexed features as the fusion features, which are used as the input of the group-relational module for grouping prediction. The group-relational module first uses the fusion features to extract features of the anchor group (i.e., nose) and predict the landmarks of the nose. It then uses the features of the anchor group and the fusion features to deduce the features of other groups and predict their landmarks. This can make full use of the structural relationship between facial components, so that the network can predict the landmarks in the case of severe occlusion.

Specifically, the three convolutional layers extract shape-indexed features F based on local patches of 34×34 size that are cropped surrounding initial landmarks from the image of 384×456 size. These shape-indexed features are first mapped into the whole features f_g and then predict a global offset value as follows:

$$\begin{aligned} f_g &= \mathbf{tanh}(W_g^1 * F + b_g^1) \\ \Delta y^g &= W_g^2 * f_g + b_g^2, \end{aligned} \tag{6}$$

where Δy^g denotes the global offset value, **tanh** denotes a nonlinear activation function, W_g^1 denotes the weights of the input-to-hidden fully connected layers, and b_g^1 denotes the biases. W_g^2 and b_g^2 are the weights and biases of the prediction layer. In the group-relational module, we introduce the shape-indexed features F into each facial group to supplement the contextual information when occlusion occurs, and the whole features f_g are also introduced to provide the global shape constraint. For the anchor group (i.e., nose), its offset value can be formulated as follows:

$$\begin{aligned} f_n &= \mathbf{tanh}(W_n^1 * (F \oplus f_g) + b_n^1) \\ \Delta y^n &= W_n^2 * f_n + b_n^2, \end{aligned} \tag{7}$$

where f_n denotes the nose group features and Δy^n denotes the offset value of all landmarks in the nose group. Even if the nose is partially occluded by other objects, shape-indexed features and global shape constraint still have the capacity to reason the robust nose feature. For other groups, in addition to the global constraints, the nose feature f_n is also introduced for relation reasoning. Taking the left group as an example, the offset of the left group is formulated as follows:

$$\begin{aligned} f_l &= \mathbf{tanh}(W_l^1 * (F \oplus f_g \oplus f_n) + b_l^1) \\ \Delta y^l &= W_l^2 * (f_l \oplus f_n) + b_l^2. \end{aligned} \tag{8}$$

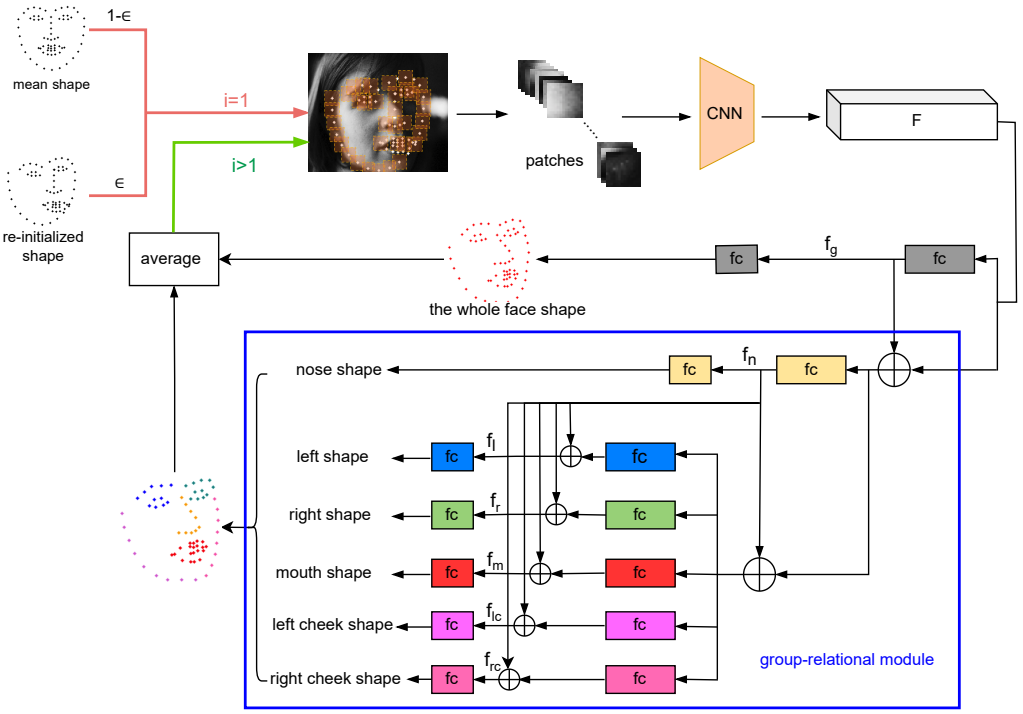

Figure 5. Illustration of a dense landmark detection network that exploits a group-relational module to reason the relations between facial groups. The mean shape and the reinitialized shape are used as candidates for initialization. $\oplus$ is the concatenation operation.

All group offsets are combined into an overall offset:

$$\Delta y^o = \Delta y^n \bigcup \Delta y^l \bigcup \Delta y^r \bigcup \Delta y^m \bigcup \Delta y^{lc} \bigcup \Delta y^{rc}, \tag{9}$$

where Δy^o denotes the prediction of the group-relational module, $\bigcup$ denotes the operation of concatenating all the offsets, $\Delta y^n, \Delta y^l, \Delta y^r, \Delta y^m, \Delta y^{lc}$, and Δy^{rc} denote the offset of the nose group, the left group, the right group, the mouth group, the left cheek group and the right cheek group, respectively.

The averaged offset value of Δy^g and Δy^o is output as the current result:

$$\Delta y = \frac{1}{2}(\Delta y^o + \Delta y^g), \tag{10}$$

and the dense landmark detection network is iterated to output the prediction:

$$y_i = y_{i-1} + \Delta y_i, \tag{11}$$

where y_i represents the coordinate of all landmarks at the i-th iteration. Mathematically, the network parameters are updated by minimizing the following objective function:

$$\min \sum_{i=1}^{I} \|y^* - (y_{i-1} + \triangle y_i)\|_2^2, \tag{12}$$

where I, y^*, y_{i-1} and $\triangle y_i$ denote the maximum iteration number, which is set to 3, the ground-truth landmarks, the results of the previous iteration, and the offset at the i-th iteration, respectively.

4. Experimentation

4.1. Datasets and Evaluation Metrics

The performance of the proposed framework STDN was validated on three datasets: 300 W [42], COFW68 [43] and Masked 300 W [13].

300 Faces In-the-Wild Challenge (300 W): This dataset [42] includes a total of 3837 faces. Each face is annotated with 68 landmarks. In our experiments, 3148 images are used as the training set, which are from the training set of LFPW and HELEN and the whole AFW. We investigate our approach by following the widely used evaluation setting: the LFPW and HELEN testing set as Commonset (554), the IBUG dataset as Challengingset (135), and the union of them as Fullset (689).

Caltech Occluded Faces in the Wild (COFW68): As proposed in [16], the COFW dataset collects faces under various occlusions and large pose variations in real life. It contains 1852 images. Each face is annotated with 29 landmarks. In our experiment, the re-annotated testing set [43] with 68 landmarks is used to verify the effectiveness of dealing with occlusion.

Masked 300 W: Masked 300 W is proposed in [13], which focuses on masked faces. It is generated by directly wearing a mask on each face. To further verify the robustness of the proposed STDN on the severely occluded face, the experiments are conducted with a cross-dataset setting: trained on the training set of 300 W and tested on three subsets of Masked 300 W.

Evaluation Metrics

Normalized mean error (NME), the curve of cumulative error distribution (CED) and the failure rate (FR) were used as evaluation metrics. The NME is defined as follows:

$$NME = \frac{1}{K} \sum_{k=1}^{K} \frac{\|S_k - S_k^*\|_2^2}{L * \Omega_k} \tag{13}$$

where S_k and S_k^* denote the predicted shape and ground truth shape, K denote the number of samples in a test set, and L, Ω_k denotes the landmark number of each face and the inter-ocular distance, respectively. The CED describes the proportion of predicted data that falls below a certain NME threshold. FR is calculated as the proportion of samples with a mean error greater than a given threshold to all samples tested, it is defined as follows:

$$FR = \frac{N_{e>e_0}}{N} \tag{14}$$

where e_0 is the set threshold, which was set to 0.1 in the experiment, $N_{e>e_0}$ represents the samples for which the normalized mean error is greater than the threshold, and N represents all samples participating in the test.

4.2. Implementation Details

In the patch resampling stage, a lightweight network is trained to detect 6 landmarks using 100×100 local patches. It starts from two convolutional layers with a stride of 1 and 32 channels for feature extraction. The convolution kernel sizes are 7×7 and 3×3, and

each layer is followed by a max-pooling operation. After that, two fully connected layers with 512-D and 12-D map features into landmark values, where 12-D is the dimensionality of landmark coordinate values. The detailed descriptions of the architectures, including the input and output shapes of each layer and the kernel sizes, are shown in Table 1. For this lightweight network, the following hyperparameters were set: an initial learning rate of 0.001, a decay factor of 0.1 and a batch size of 64.

Table 1. Architecture of the lightweight network used in the patch resampling stage.

Layers	Input Shape	Output Shape	Kernel
conv1	$6 \times 100 \times 100 \times 3$	$6 \times 94 \times 94 \times 32$	$[7 \times 7, 32]$
pool1	$6 \times 94 \times 94 \times 32$	$6 \times 47 \times 47 \times 32$	$[2 \times 2, 32]$
conv2	$6 \times 47 \times 47 \times 32$	$6 \times 45 \times 45 \times 32$	$[3 \times 3, 32]$
pool2	$6 \times 45 \times 45 \times 32$	$6 \times 22 \times 22 \times 32$	$[2 \times 2, 32]$
fc1	$6 \times 22 \times 22 \times 64$	1×512	-
fc2	1×512	6×2	-

Next, in the relation reasoning stage, fine detection of landmarks is carried out, in which three convolutional layers (7×7, 3×3, 3×3 kernel size) are used to extract shape-indexed features from 34×34 size local patches. Each convolutional layer with a stride of 1 and 32 channels is followed by a max-pooling layer. After that, the features are fed into the group-relational module. Each branch network uses two fully connected layers with 512-D and d, where the value of d changes according to the number of landmarks in the corresponding group. The detailed descriptions of the architectures, including the input and output shapes of each layer and the kernel sizes, are shown in Table 2. For this network, an initial learning rate of 0.0002 and a decay factor of 0.97 were used.

Table 2. Architecture of the dense landmark detection network used in the relation reasoning stage. The value of d changes according to the number of landmarks in the group.

Layers	Input Shape	Output Shape	Kernel
conv1	$68 \times 34 \times 34 \times 3$	$68 \times 28 \times 28 \times 32$	$[7 \times 7, 32]$
pool1	$68 \times 28 \times 28 \times 32$	$68 \times 14 \times 14 \times 32$	$[2 \times 2, 32]$
conv2	$68 \times 14 \times 14 \times 32$	$68 \times 12 \times 12 \times 32$	$[3 \times 3, 32]$
pool2	$68 \times 12 \times 12 \times 32$	$68 \times 6 \times 6 \times 32$	$[2 \times 2, 32]$
conv3	$68 \times 6 \times 6 \times 32$	$68 \times 4 \times 4 \times 32$	$[3 \times 3, 32]$
pool3	$68 \times 4 \times 4 \times 32$	$68 \times 2 \times 2 \times 32$	$[2 \times 2, 32]$
fc1	$68 \times 2 \times 2 \times 64$	1×512	-
fc2	1×512	68×2 *or* d	-

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

4.3.1. Evaluation on the 300 W Dataset

Table 3 shows the comparisons of the sparse-to-dense network (STDN) with state-of-the-art methods on 300 W [42] dataset. Compared with the same type of coordinate regression methods [12,13,18,21,34,41,44,45], STDN significantly outperforms other methods on Commonset, Challengingset, and Fullset. A recent coordinate regression method, SRN [13], which has comparable performance, can solve occluded faces by exploring the spatial dependence between different facial components over long and short distances. However, it does not perform well on faces with large pose variations. RMTL [45] focuses on using the complementary information between facial landmark localization and expression recognition to improve performance, and the proposed residual learning module enables the two tasks to learn complementary information from each other. Our method STDN focuses on obtaining enough information from face images to locate landmarks without multitasking. The proposed STDN achieves an NME value of 5.33% on the Challengingset.

Table 3. NME (in %) of 68-point landmark detection on 300 W.

	Method	Challenging	Common	Full
Heatmap regression	LAB [11] (2018)	5.19	2.98	3.41
	SAN [26] (2018)	6.60	3.34	3.98
	DFL [28] (2019)	7.20	4.11	4.72
	RWAN [30] (2019)	7.37	3.21	3.97
	ADC [31] (2020)	7.04	2.83	4.23
	3FabRec [32] (2020)	5.74	3.36	3.82
	MTAAE [37] (2021)	7.48	4.30	5.30
Coordinate regression	MDM[12] (2016)	7.48	4.03	4.46
	DR [18] (2016)	13.80	4.51	6.31
	TR-DRN [21] (2017)	7.56	4.36	4.99
	ODN [34] (2019)	6.67	3.56	4.17
	AVS [44] (2019)	6.49	3.21	3.86
	SRN [13] (2021)	5.86	3.08	3.62
	RMTL [45] (2021)	5.50	**3.00**	3.49
	mnv2 [41] (2022)	6.13	3.56	4.06
	STDN (Ours)	**5.33**	3.02	**3.44**

In comparison with the heatmap regression methods [11,26,28,30–32,37], the proposed STDN can still exceed most of the methods on three subsets. However, heatmap regression methods always stack deeper networks, so they are slower to compute and require a larger number of parameters. For example, in terms of inference speed, our method can achieve 31 FPS, outperforming LAB [11] (11 FPS) by a large margin.

Figure 6 shows the CED curves of STDN compared to the methods [12,46–51]. It can be seen that STDN significantly outperforms these open-source face alignment methods. Figure 7 shows the qualitative results of STDN on 300 W [42]. Under large pose variations and occlusion, the predictions from STDN still ensure the overall face shape.

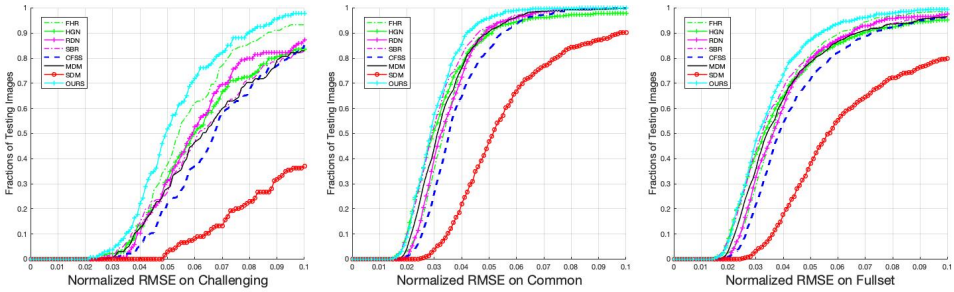

Figure 6. The CED curves for three subsets of 300 W.

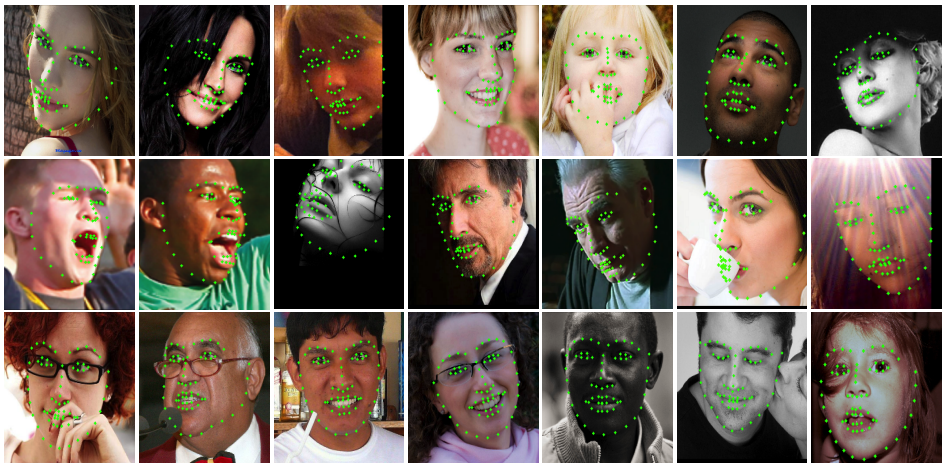

Figure 7. Representative results on 300 W.

4.3.2. Evaluation on COFW68 Dataset

To further prove the robustness of STDN when handling occluded faces, we conducted a cross-dataset evaluation on COFW68 [43] dataset, which covers different occlusions. In this setting, COFW68 is only used for testing, not training. Table 4 shows the performance compared with other methods on COFW68. We find that only LAB [11] outperforms ours in terms of NME, but STDN achieves the lowest failure rate. LAB [11] is a heatmap-based method, which is computationally expensive. The qualitative results of COFW68 are visualized in Figure 8.

Table 4. NME (in %) of 68-point landmark detection on COFW68.

Method	NME	FR (0.1)
HPM [43] (2014)	7.46	-
OSRD [52] (2014)	9.27	-
TCDCN [53] (2015)	8.05	6.31
LBF [54] (2016)	13.7	-
CRASM [19] (2016)	8.02	-
MDM [12] (2016)	6.32	4.31
LAB [11] (2018)	**4.62**	2.17
ODN [34] (2019)	5.87	2.84
STDN (Ours)	4.64	**1.97**

Figure 8. Representative results on COFW68.

4.3.3. Evaluation on Masked 300 W Dataset

Although COFW68 [43] and the Challengingset of 300 W [42] contain a large number of real-life occlusions, these occlusions are often small, and severe occlusions, such as medical masks, are rarely seen. Table 5 shows the comparison results with state-of-the-art methods [12,46,49,50,55,56] on Masked 300 W [13]. The proposed STDN achieves the best performance on three subsets. Figure 9 displays the visualized predictions of STDN on Masked 300 W [13]. Our predictions demonstrate that STDN can reason for rational face structures even under severe occlusions.

Table 5. NME (in %) of 68-point landmark detection on Masked 300 W.

Method	Challenging	Common	Full
CFSS [46] (2015)	19.98	11.73	13.35
MDM [12] (2016)	15.66	8.42	9.83
SHG [49] (2016)	13.52	8.17	9.22
FAN [55] (2017)	10.81	7.36	8.02
SBR [50] (2018)	15.28	9.72	10.65
DHGN [56] (2020)	12.19	8.98	9.61
STDN (Ours)	**9.64**	**6.51**	**7.12**

Figure 9. Example images from Masked 300 W.

4.3.4. Analysis

From the above results, it can be seen that STDN demonstrates competitive performance compared with state-of-the-art approaches. This is mainly due to three advantages: (1) Compared with existing methods based on shape-indexed features, STDN obtains a high-quality reinitialized shape that can be used to crop high-quality local patches; (2) the shape constraints implied in shape-indexed features may fail under severe occlusion, and GPR further explores the spatial relationships between facial groups to strengthen the shape constraints; and (3) compared with LAB [11], which imposes shape constraints by predicting facial boundary information, STDN has a faster inference speed.

4.4. Ablation Study

4.4.1. Investigation of the Effectiveness of the Two Stages

To investigate the impact of the patch resampling stage (PRS) and the group-relational module (GRM) on landmark detection, four experiments were carried out on 300 W [42]: (1) PRS and GRM were removed from STDN as baseline, the prediction was obtained by two fully connected layers; (2) PRS was added to the baseline; (3) GRM was added to the baseline, the shape-indexed patches were cropped by using a mean shape; and (4) PRS and GRM were added to the baseline. The results are shown in Table 6. Both PRS and GRM can achieve improvements in detection accuracy. The best performance is achieved when integrating both PRS and GRM into the STDN.

Table 6. Ablation experiments of the patch resampling stage (PRS) and the group-relational module (GRM).

Method	Challenging	Common	Full
baseline	6.31	3.23	3.82
baseline + PRS	5.55	3.07	3.55
baseline + GRM	5.58	3.05	3.54
baseline + PRS + GRM	5.33	3.02	3.44

4.4.2. Investigation of Different Group Strategies

We further investigated the different strategies of facial group division. To report the experiments, four group divisions were exploited, as shown in Figure 10. GRM* divides the whole face into three groups: the upper group includes eyes and eyebrows, the lower group includes nose and mouth and the cheek group; GRM** divides the whole face into five groups: the eyebrows group, the eyes group, the nose group, the mouth group and the cheek group; GRM*** divides the whole face into four groups: the upper group includes eyebrows and eyes, the nose group,the mouth group and the cheek group; GRM is a face division method used in STDN. These divisions explore the geometric relations

between facial components. The results are reported in Table 7, which indicates that the best performance can be achieved by dividing the face into five groups with the nose as the center.

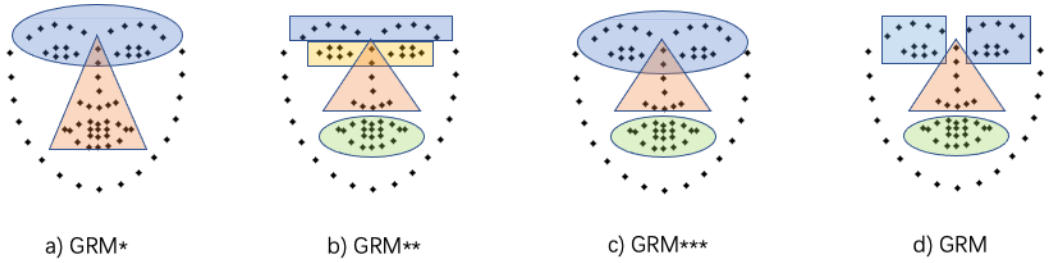

Figure 10. Strategies for dividing face shape into different facial groups.

Table 7. Ablation experiments of different group division strategies on 300 W. *, **, *** indicate the results using the different group division strategies as shown in Figure 10.

Method	Challenging	Common	Full
STDN with GRM*	5.57	3.09	3.57
STDN with GRM**	5.53	3.06	3.54
STDN with GRM***	5.48	3.04	3.51
STDN with GRM	5.33	3.02	3.44

4.4.3. Investigation of Hyperparameters

In the relation reasoning stage, the hyperparameter ϵ is introduced to avoid overfitting. We increased the value of ϵ from 0 to 1 in steps of 0.1 and reported its role in localization accuracy in Table 8. The results show that STDN achieves the best performance when ϵ is set to 0.5.

Table 8. Ablation experiments of hyperparameter ϵ on 300 W Challengingset.

ϵ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
NME	7.28	5.41	5.40	5.41	5.38	**5.33**	5.41	5.45	5.46	5.59	5.54

5. Conclusions

This paper proposes a sparse-to-dense network (STDN) to deal with occlusion problems. The proposed framework employs a patch resampling approach to improve the quality of shape-indexed patches, which is helpful for extracting robust shape-indexed features. Moreover, the STDN exploits the group relations between facial components to handle occluded faces using a carefully designed group-relational module. Extensive experiments were conducted to evaluate the performance of the STDN in normal conditions and occlusion. The experimental results show that STDN improves by 9.16% on Fullset compared to the baseline, achieves 5.33% on the Challengingset, and improves by 15.53% compared to the baseline, which fully demonstrates that STDN outperforms most current methods in terms of robustness against occlusion. Currently, STDN only considers the nose as an anchor to constrain other facial groups, future work can learn occlusion-adaptive group relations to make full use of the spatial relations of faces, and also consider learning differential loss function, which aims at adaptively focusing on the occluded region.

Author Contributions: X.W. made contributions to conception and manuscript writing; X.L. examined and supervised this research and outcomes; Y.W. revised and polished the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: In this manuscript, the employed datasets have been taken with license agreements from the corresponding institutions with proper channels.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xiong, F.; Xiao, Y.; Cao, Z.; Wang, Y.; Zhou, J.T.; Wu, J. Ecml: An ensemble cascade metric-learning mechanism toward face verification. *IEEE Trans. Cybern.* **2022**, *52*, 1736–1749. [CrossRef] [PubMed]
2. Saoud, A.; Oumane, A.; Ouafi, A.; Taleb-Ahmed, A. Multimodal 2d+ 3d multi-descriptor tensor for face verification. *Multimed. Tools Appl.* **2020**, *79*, 23071–23092. [CrossRef]
3. Singh, M.; Nagpal, S.; Singh, R.; Vatsa, M. Disguise resilient face verification. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 3895–3905. [CrossRef]
4. Zhang, H.; Su, W.; Yu, J.; Wang, Z. Identity–expression dual branch network for facial expression recognition. *IEEE Trans. Cogn. Dev. Syst.* **2020**, *13*, 898–911. [CrossRef]
5. Zhang, Z.; Lai, C.; Liu, H.; Li, Y.-F. Infrared facial expression recognition via gaussian-based label distribution learning in the dark illumination environment for human emotion detection. *Neurocomputing* **2020**, *409*, 341–350. [CrossRef]
6. Gaddam, D.K.R.; Ansari, M.D.; Vuppala, S.; Gunjan, V.K.; Sati, M.M. Human facial emotion detection using deep learning. In *ICDSMLA 2020*; Springer: Berlin, Germany, 2022; pp. 1417–1427.
7. Liu, K.; Cao, G.; Zhou, F.; Liu, B.; Duan, J.; Qiu, G. Towards disentangling latent space for unsupervised semantic face editing. *IEEE Trans. Image Process.* **2022**, 1475–1489. [CrossRef]
8. Hou, X.; Zhang, X.; Liang, H.; Shen, L.; Lai, Z.; Wan, J. Guidedstyle: Attribute knowledge guided style manipulation for semantic face editing. *Neural Netw.* **2022**, *145*, 209–220. [CrossRef]
9. Lu, Y.; Khan, M.; Ansari, M.D. Face recognition algorithm based on stack denoising and self-encoding lbp. *J. Intell. Syst.* **2022**, *31*, 501–510. [CrossRef]
10. Talab, M.A.; Awang, S.; Ansari, M.D. A novel statistical feature analysis-based global and local method for face recognition. *Int. J. Opt.* **2020**, *2020*, 4967034. [CrossRef]
11. Wu, W.; Qian, C.; Yang, S.; Wang, Q.; Cai, Y.; Zhou, Q. Look at boundary: A boundary-aware face alignment algorithm. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2129–2138.
12. Trigeorgis, G.; Snape, P.; Nicolaou, M.A.; Antonakos, E.; Zafeiriou, S. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4177–4187.
13. Zhu, C.; Li, X.; Li, J.; Dai, S.; Tong, W. Reasoning structural relation for occlusion-robust facial landmark localization. *Pattern Recognit.* **2021**, *122*, 108325. [CrossRef]
14. Cootes, T.F.; Taylor, C.J.; Cooper, D.H.; Graham, J. Active shape models-their training and application. *Comput. Vis. Image Underst.* **1995**, *61*, 38–59. [CrossRef]
15. Cootes, T.F.; Edwards, G.J.; Taylor, C.J. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 681–685. [CrossRef]
16. Burgos-Artizzu, X.P.; Perona, P.; Dollár, P. Robust face landmark estimation under occlusion. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1513–1520.
17. Huang, Z.; Zhou, E.; Cao, Z. Coarse-to-fine face alignment with multi-scale local patch regression. *arXiv* **2015**, arXiv:1511.04901.
18. Shi, B.; Bai, X.; Liu, W.; Wang, J. Face alignment with deep regression. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *29*, 183–194. [CrossRef]
19. Liu, Q.; Deng, J.; Yang, J.; Liu, G.; Tao, D. Adaptive cascade regression model for robust face alignment. *IEEE Trans. Image Process.* **2016**, *26*, 797–807. [CrossRef]
20. Park, B.-H.; Oh, S.-Y.; Kim, I.-J. Face alignment using a deep neural network with local feature learning and recurrent regression. *Expert Syst. Appl.* **2017**, *89*, 66–80. [CrossRef]
21. Lv, J.; Shao, X.; Xing, J.; Cheng, C.; Zhou, X. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3317–3326.
22. Feng, Z.-H.; Kittler, J.; Christmas, W.; Huber, P.; Wu, X.-J. Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2481–2490.
23. Cao, X.; Wei, Y.; Wen, F.; Sun, J. Face alignment by explicit shape regression. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2887–2894.
24. Wu, Y.; Ji, Q. Shape augmented regression method for face alignment. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 26–32.

25. Xiao, S.; Feng, J.; Xing, J.; Lai, H.; Yan, S.; Kassim, A. Robust facial landmark detection via recurrent attentive-refinement networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin, Germany, 2016; pp. 57–72.
26. Dong, X.; Yan, Y.; Ouyang, W.; Yang, Y. Style aggregated network for facial landmark detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 379–388.
27. Tang, Z.; Peng, X.; Geng, S.; Wu, L.; Zhang, S.; Metaxas, D. Quantized densely connected u-nets for efficient landmark localization. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 339–354.
28. Chen, W.; Zhou, Q.; Hu, H. Face alignment by discriminative feature learning. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 2204–2208.
29. Wang, X.; Bo, L.; Fuxin, L. Adaptive wing loss for robust face alignment via heatmap regression. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6971–6981.
30. Wang, L.; Xiang, W. Residual neural network and wing loss for face alignment network. In Proceedings of the 2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Dalian, China, 14–16 November 2019; pp. 1093–1097.
31. Chandran, P.; Bradley, D.; Gross, M.; Beeler, T. Attention-driven cropping for very high resolution facial landmark detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5861–5870.
32. Browatzki, B.; Wallraven, C. 3fabrec: Fast few-shot face alignment by reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 6110–6120.
33. Kumar, A.; Chellappa, R. Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 430–439.
34. Zhu, M.; Shi, D.; Zheng, M.; Sadiq, M. Robust facial landmark detection via occlusion-adaptive deep networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3486–3496.
35. Kumar, A.; Marks, T.K.; Mou, W.; Wang, Y.; Jones, M.; Cherian, A.; Koike-Akino, T.; Liu, X.; Feng, C. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 8236–8246.
36. Wan, J.; Lai, Z.; Shen, L.; Zhou, J.; Gao, C.; Xiao, G.; Hou, X. Robust facial landmark detection by cross-order cross-semantic deep network. *Neural Netw.* **2021**, *136*, 233–243. [CrossRef]
37. Yue, X.; Li, J.; Wu, J.; Chang, J.; Wan, J.; Ma, J. Multi-task adversarial autoencoder network for face alignment in the wild. *Neurocomputing* **2021**, *437*, 261–273. [CrossRef]
38. Zhu, C.; Li, X.; Li, J.; Dai, S. Improving robustness of facial landmark detection by defending against adversarial attacks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11751–11760.
39. Ma, J.; Li, J.; Du, B.; Wu, J.; Wan, J.; Xiao, Y. Robust face alignment by dual-attentional spatial-aware capsule networks. *Pattern Recognit.* **2022**, *122*, 108297. [CrossRef]
40. Wang, H.; Cheng, R.; Zhou, J.; Tao, L.; Kwan, H.K. Multistage model for robust face alignment using deep neural networks. *Cogn. Comput.* **2022**, *14*, 1123–1139. [CrossRef]
41. Fard, A.P.; Mahoor, M.H. Facial landmark points detection using knowledge distillation-based neural networks. *Comput. Vis. Image Underst.* **2022**, *215*, 103316. [CrossRef]
42. Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; Pantic, M. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Sydney, Australia, 2–8 December 2013; pp. 397–403.
43. Ghiasi, G.; Fowlkes, C.C. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2385–2392.
44. Qian, S.; Sun, K.; Wu, W.; Qian, C.; Jia, J. Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 10153–10163.
45. Chen, B.; Guan, W.; Li, P.; Ikeda, N.; Hirasawa, K.; Lu, H. Residual multi-task learning for facial landmark localization and expression recognition. *Pattern Recognit.* **2021**, *115*, 107893. [CrossRef]
46. Zhu, S.; Li, C.; Loy, C.C.; Tang, X. Face alignment by coarse-to-fine shape searching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4998–5006.
47. Liu, H.; Lu, J.; Guo, M.; Wu, S.; Zhou, J. Learning reasoning-decision networks for robust face alignment. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 679–693. [CrossRef] [PubMed]
48. Xiong, X.; De la Torre, F. Supervised descent method and its applications to face alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 532–539.
49. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin, Germany, 2016; pp. 483–499.

50. Dong, X.; Yu, S.-I.; Weng, X.; Wei, S.-E.; Yang, Y.; Sheikh, Y. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 360–368.
51. Tai, Y.; Liang, Y.; Liu, X.; Duan, L.; Li, J.; Wang, C.; Huang, F.; Chen, Y. Towards highly accurate and stable face alignment for high-resolution videos. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8893–8900.
52. Xing, J.; Niu, Z.; Huang, J.; Hu, W.; Yan, S. Towards multi-view and partially-occluded face alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1829–1836.
53. Zhang, Z.; Luo, P.; Loy, C.C.; Tang, X. Learning deep representation for face alignment with auxiliary attributes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 918–930. [CrossRef]
54. Ren, S.; Cao, X.; Wei, Y.; Sun, J. Face alignment via regressing local binary features. *IEEE Trans. Image Process.* **2016**, *25*, 1233–1245. [CrossRef] [PubMed]
55. Bulat, A.; Tzimiropoulos, G. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1021–1030.
56. Zhu, H.; Liu, H.; Zhu, C.; Deng, Z.; Sun, X. Learning spatial-temporal deformable networks for unconstrained face alignment and tracking in videos. *Pattern Recognit.* **2020**, *107*, 107354. [CrossRef]