mathematics

MDPI

Article

Region-Aware Deep Feature-Fused Network for Robust Facial Landmark Localization

Xuxin Lin and Yanyan Liang *

Faculty of Innovation Engineering, Macau University of Science and Technology, Macau 999078, China; xxlin@must.edu.mo

* Correspondence: yyliang@must.edu.mo

Abstract: In facial landmark localization, facial region initialization usually plays an important role in guiding the model to learn critical face features. Most facial landmark detectors assume a well-cropped face as input and may underperform in real applications if the input is unexpected. To alleviate this problem, we present a region-aware deep feature-fused network (RDFN). The RDFN consists of a region detection subnetwork and a region-wise landmark localization subnetwork to explicitly solve the input initialization problem and derive the landmark score maps, respectively. To exploit the association between tasks, we develop a cross-task feature fusion scheme to extract multi-semantic region features while trading off their importance in different dimensions via global channel attention and global spatial attention. Furthermore, we design a within-task feature fusion scheme to capture the multi-scale context and improve the gradient flow for the landmark localization subnetwork. At the inference stage, a location reweighting strategy is employed to transform the score maps into 2D landmark coordinates. Extensive experimental results demonstrate that our method has competitive performance compared to recent state-of-the-art methods, achieving NMEs of 3.28%, 1.48%, and 3.43% on the 300W, AFLW, and COFW datasets, respectively.

Keywords: facial landmark localization; face alignment; region-based CNN; deep feature fusion

MSC: 68T07

Citation: Lin, X.; Liang, Y. Region-Aware Deep Feature-Fused Network for Robust Facial Landmark Localization. *Mathematics* **2023**, *11*, 4026. https://doi.org/10.3390/math11194026

Academic Editors: Costin Badica, Nick Bassiliades, Kalliopi Kravari and Theodoros Kosmanis

Received: 21 August 2023
Revised: 13 September 2023
Accepted: 15 September 2023
Published: 22 September 2023

1. Introduction

Facial landmark localization, also known as face alignment, aims to detect a set of semantic points on a face image, including eye corners, mouth corners, nose tip, etc. It typically serves as a critical step in many computer vision applications, such as face recognition, face attribute analysis, and human–computer interaction. Over the past decades, many studies have been reported in the literature to improve the performance of landmark localization. However, it remains a challenging problem to develop a robust facial landmark detector that accurately detects different landmarks in unconstrained face images with illumination variations, large poses, and strong occlusion.

In previous works, the methods based on conventional cascaded shape regression (CSR) [1–3] made significant progress in facial landmark localization by directly learning a mapping function to iteratively correct the landmark position estimation. Although these methods have been successful in solving near-frontal face alignment, their accuracy is dramatically degraded on some challenging datasets, such as 300W [4], AFLW [5], and COFW [2]. One of the main reasons is that the extracted image features are either hand-crafted features, such as SIFT [1] or simply learned features [6], which are weakly discriminative for unconstrained face images.

Recently, with the advancement of deep learning techniques in computer vision, deep neural network (DNN)-based methods [7–9] have demonstrated superior performance on the challenging benchmarks. These methods typically assume that a well-cropped

facial region is used for input initialization. However, they may underperform in real-world applications if an off-the-shelf face detector cannot provide a suitable bounding box to capture the expected facial region. Some recent works [10–12] have focused on solving the initialization problem by cascading multiple DNNs to explicitly reinitialize the input image. In our work, we propose a new deep network architecture with automatic region initialization. As shown in Figure 1, given a coarse face image as input, the critical facial region is first detected by a region detection subnetwork. Then, the corresponding region features are extracted and fed into a landmark localization subnetwork to obtain the landmark score maps. Finally, these score maps are transformed into a set of 2D landmark coordinates using a location reweighting strategy. In contrast to the previous works, a key difference is that we concatenate all the subnetworks into an end-to-end architecture by designing a cross-task feature fusion scheme and a within-task feature fusion scheme. The contributions of our work are described as follows:

- We present an end-to-end deep convolutional network called region-aware deep feature-fused network (RDFN). The proposed network can simultaneously solve the region initialization problem and the facial landmark localization task.
- In the RDFN, we design two efficient feature fusion schemes to derive the cross-task and within-task feature representations, which further improve the accuracy of facial landmark localization on various unconstrained face images.
- At the inference stage, we introduce a location reweighting strategy to effectively transform the landmark score maps into 2D landmark coordinates.
- We perform extensive experiments to demonstrate the effectiveness of the proposed components and the superior performance of our approach on several challenging datasets, including 300W, AFLW, and COFW.

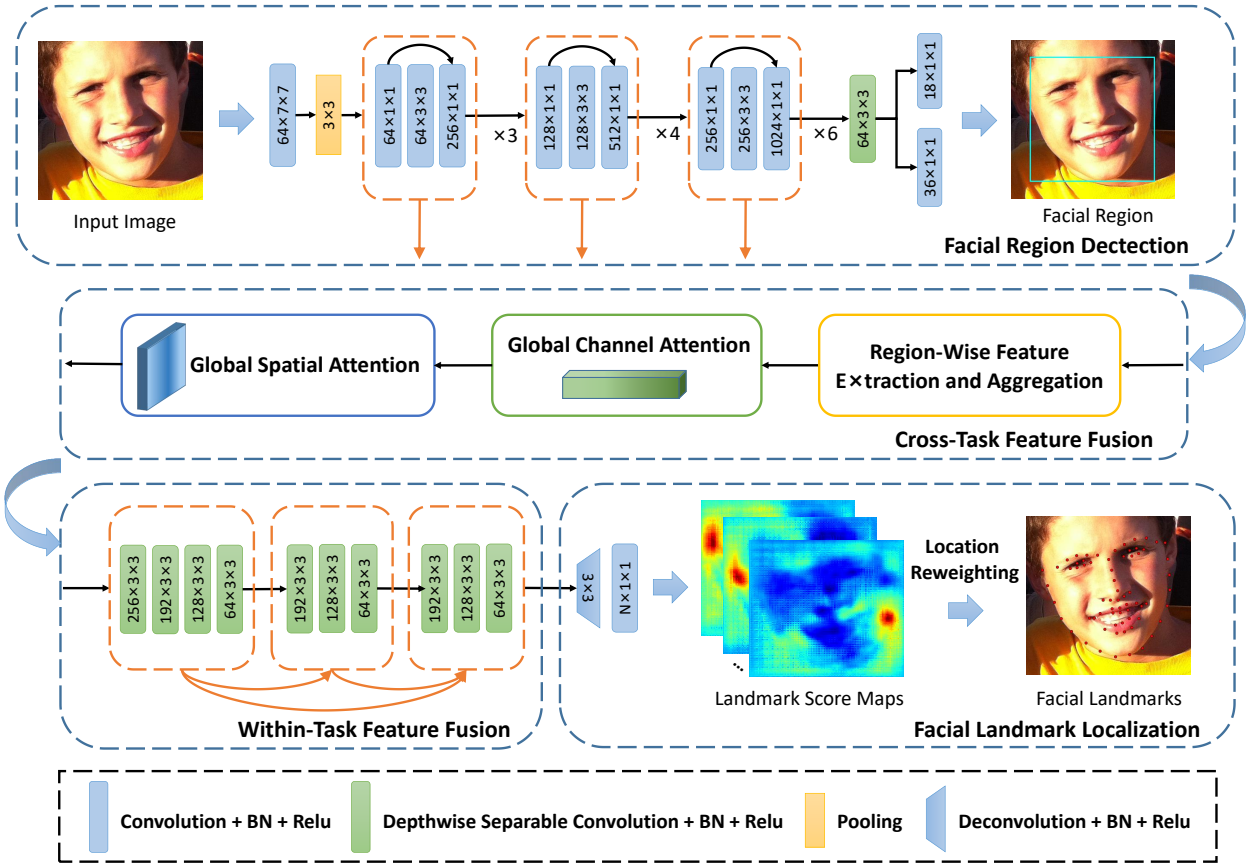

Figure 1. Overview of the proposed region-aware deep feature-fused network for facial landmark localization. The landmark score map is a heatmap that ranges from blue to red and represents the spatial probability distribution of a landmark from 0 to 1.

2. Related Work

In the early literature of the last decade, the conventional CSR-based methods represent a significant milestone in facial landmark localization. They formulate the landmark localization task as a regression problem, which can be solved by creating a mapping function from extracted image features to estimated landmark positions. The key to the success of CSR-based methods is the use of cascaded weak regressors to iteratively refine the landmark position estimation. Several classical approaches have been proposed using different regression methods and extracted features. The supervised descent method (SDM) [1] uses the scale-invariant feature transform (SIFT) features and constructs a cascaded linear regressor for facial landmark prediction. Robust cascaded pose regression (RCPR) [2] and explicit shape regression (ESR) [13] apply the shape-indexed features to a multi-stage boosted regression method to estimate the landmark coordinates. Local binary features (LBFs) [6] accelerate the landmark estimation process using the local binary features and a classical regression random forest to learn the mapping function. Cascaded collaborative regression (CCR) [14] combines the dynamic multi-scale histograms of oriented gradient (HOG) features with a cascaded linear regression method for facial landmark detection. Coarse-to-fine shape searching (CFSS) [15] improves the CSR regression scheme using the SIFT features and the binary robust independent elementary (BRIEF) features in a coarse-to-fine shape searching method.

In recent years, deep learning methods for face landmark localization have received more attention than conventional CSR-based methods because DNN can learn and extract more discriminative features from face images. The deep learning methods can be simply divided into the DNN-based method without or with reinitialization. The method without reinitialization usually takes a well-cropped face image as input to the DNN and does not introduce the refinement operations on the raw input. The mnemonic descent method (MDM) [16] and recurrent attentive refinement (RAR) [17] propose an end-to-end recurrent convolutional network for coarse-to-fine facial landmark refinement. Task-constrained deep convolutional network (TCDCN) [18] is a multi-task CNN architecture for jointly learning the tasks of face alignment and facial attribute analysis. With the development of large DNN models, some works use more complex network architectures to enhance the feature extraction capability, such as the high-resolution representation network (HRNet) [19] and stacked hourglass network [20] with adaptive wing loss (AWing) [7]. To achieve efficient landmark localization, some recent works [8,21] have started to explore how to reduce the computational cost using lightweight DNN models.

Although these methods have achieved impressive performance on various challenging benchmarks, they may underperform in real-world applications without an appropriate facial region as input. To alleviate this problem, the DNN-based method with reinitialization has been proposed to refine the raw input in the inference process. In the early works, most of the methods [10,22,23] cascade multiple convolutional networks to perform coarse-to-fine landmark estimation by iteratively reinitializing the input image of the global or local facial region at each stage. To reduce the number of cascaded networks, PicassoNet [9] proposes a single network with group convolution, which uses each convolution group to predict the landmarks of each local facial region instead of using each individual network. Recently, some works [11,12] tend to construct an end-to-end DNN architecture using the refined facial region or facial parts to reinitialize the input features.

From the above literature, we can see that deep learning methods have significantly advanced the development of facial landmark localization in recent years. An increasing number of researchers aim to develop a practical facial landmark detector by explicitly reinitializing the input to reduce the dependence on the well-cropped face image. Inspired by these works, we propose a region-aware convolutional network to jointly solve the input initialization problem and the facial landmark localization task. Compared to existing DNN-based methods with reinitialization, our method is based on an end-to-end region-aware network architecture with a lightweight component design, which can effectively trade off model complexity and inference performance.

3. Methodology

3.1. Motivation and Overview

Inspired by the region-based multi-task CNN (RM-CNN) architecture for jointly solving detection and other vision tasks, such as object segmentation in Mask R-CNN [24], we find that the detection process in RM-CNN can locate the regions of interest (RoIs) used to reinitialize the input region-wise features for subsequent tasks. In contrast to the cascaded DNN methods with the multi-stage reinitialization of an input image in different networks, RM-CNN can solve the region initialization problem directly through a built-in detection subnetwork working with a task-specific subnetwork in an end-to-end manner. Nevertheless, a vanilla RM-CNN would underperform in the facial landmark localization task for the following reasons:

- The input features are only extracted from the top layer of the network backbone and are weakly discriminative for the fine-grained tasks due to the lack of low-level semantic information from the lower layers.
- In the downstream task, the ability to extract features is further limited by the simple stacked encoding structure of the task-specific subnetwork, which only considers a single gradient flow between network layers.

In our work, we aim to propose a new RM-CNN model with two feature fusion schemes to address the above problems and achieve robust facial landmark localization. As shown in Figure 1, the network backbone of our model is constructed by following the ResNet-50 [25] architecture and using the top three groups of bottleneck residual blocks with repetition numbers of 3, 4, and 6. Each residual block contains three $C \times W_k \times H_k$ convolution operations with a skip connection, where C, W_k and H_k denote the number of channels and the width and height of a kernel, respectively. Batch normalization (BN) and ReLu activation are the standard operations performed after each convolution operation in our model. Through bottom–up continuous encoding, we can obtain hierarchical feature maps with different levels of semantics. On top of the backbone module, we add a lightweight detection head consisting of a depthwise separable convolution and two decision convolutions to generate the $4 \times N_r \times W_i/16 \times H_i/16$ and $2 \times N_r \times W_i/16 \times H_i/16$ feature maps, where N_r, W_i and H_i denote the number of reference bounding boxes in each output pixel and the width and height of an input image, respectively. From the feature maps, we can derive two classes (facial region/non-facial region) and four parameterized coordinates of all reference bounding boxes.

To obtain the region-wise input features with rich semantic information, we introduce a cross-task feature fusion scheme. First, we extract and aggregate the facial features from the top block in each bottleneck residual group using the predicted region bounding boxes. Then, we trade off the importance of these features in the channel and spatial denominations using the global channel and spatial attention blocks, respectively. In the downstream subnetwork, we replace the original stacked encoding scheme with a within-task feature fusion scheme to explicitly improve the gradient flow and capture the multi-scale context. At the landmark localization stage, we use a $64 \times 3 \times 3$ deconvolution operation with a stride of 9 and a $(N_l + 1) \times 1 \times 1$ convolution operation to transform the final feature maps into $(N_l + 1)$ $W_l \times H_l$ landmark score maps, where N_l, W_l and H_l denote the number of predicted landmarks and the width and height of a landmark score map, respectively. Finally, we can obtain a set of 2D landmark coordinates by applying a location reweighting strategy to these score maps.

3.2. Cross-Task Feature Fusion Scheme

Given an RoI from the region detection subnetwork as input, we first perform a RoIAlign operation with L2 normalization to extract and concatenate a set of region-wise feature maps $F_r \in \mathbb{R}^{C_a \times W_r \times H_r}$, where C_a, W_r, and H_r denote the number of aggregated channels and the width and height of the feature maps, respectively. The RoIAlign operation is defined in the work [24] to ensure that the features are extracted from the same spatial location as the RoI, preserving spatial accuracy and avoiding misalignment. L2

normalization is applied to normalize the feature values at each spatial location across different channels, mitigating the scalar differences that may occur among the feature maps from different layers. Then, through two proposed global attention blocks as shown in Figure 2, we can obtain a channel weight vector $A_c \in \mathbb{R}^{C_a \times 1 \times 1}$ and a spatial weight map $A_s \in \mathbb{R}^{1 \times W_r \times W_r}$. The region-wise feature maps are reweighted by A_c and A_s as follows:

$$\begin{aligned} F_r^{'} &= M_{broadcast}(A_c, F_r) \otimes F_r, \\ F_r^{''} &= M_{broadcast}(A_s, F_r^{'}) \otimes F_r^{'}, \end{aligned} \tag{1}$$

where $M_{broadcast}$ is the broadcast function to adapt the weight vector or map to the shape of the region-wise feature maps by copying the values along the channel or spatial dimension. $\otimes$ denotes an element-wise multiplication operation. In the following, we detail the process of two global attention blocks.

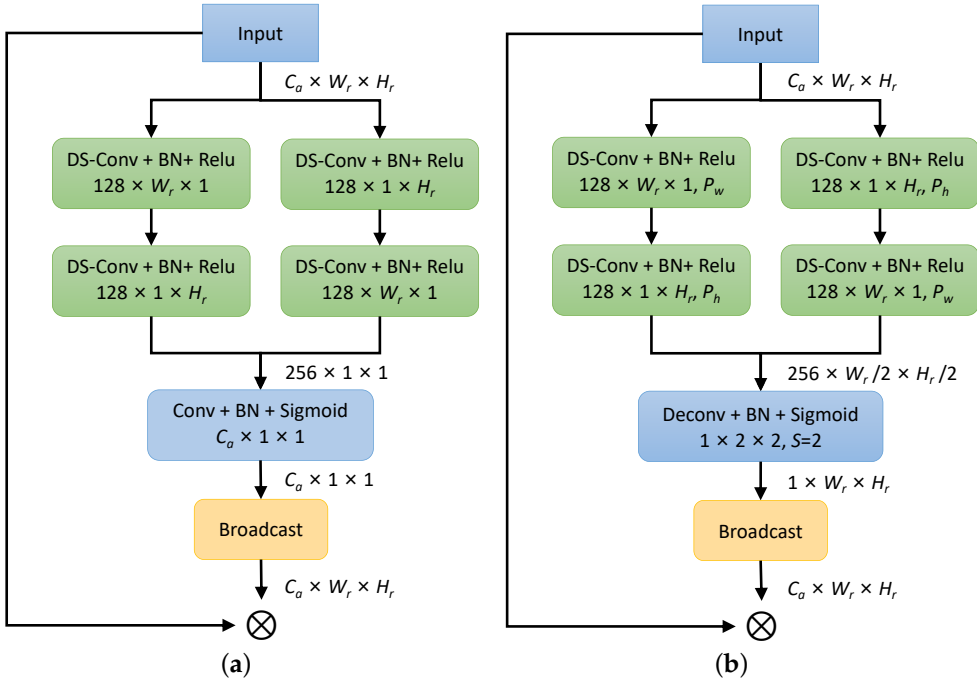

Figure 2. Overview of the proposed global channel attention block (**a**) and global spatial attention block (**b**).

3.2.1. Global Channel Attention Block

In CNN, the channel attention focuses on the channel dimension of the input feature maps. By learning the interdependencies among different channels, the channel attention can generate a channel weight vector to find which feature map is important in a latent space. This process allows the network to emphasize important channel features while suppressing irrelevant ones. The channel weight vector is given as follows:

$$A_c = Sigmoid(F_c(F_r)), \tag{2}$$

where F_c denotes a learned composite function that receives the input features and outputs a channel-wise weight vector. In the previous works [26,27], F_c usually includes a global max-pooling or average-pooling operation to squeeze the spatial information of the feature maps, and a multi-layer perceptron (MLP) with a hidden layer to learn the weight vector. This design cannot explicitly learn the potential relationship among the feature pixels in different locations for the contribution to each channel weight. An intuitive method is to use a global convolution operation with a $W_r \times H_r$ kernel instead of the pooling operation.

However, this results in a large number of parameters, i.e., ${C_a}^2 \times W_r \times H_r$ #Params, and high computational cost, i.e., ${C_a}^2 \times W_r \times H_r$ multiply–accumulate operations (MACC).

In our work, we combine the depthwise and spatially separable convolutions to efficiently achieve global feature encoding. As shown in Figure 2a, the input feature maps are separately and continuously encoded along the horizontal and vertical directions by two groups of depthwise separable convolutions (DS-Convs). Each group contains a $128 \times W_r \times 1$ DS-Conv and a $128 \times 1 \times H_r$ DS-Conv, which work together to perform feature encoding with a global receptive field. The DS-Conv consists of a depthwise convolution and a pointwise convolution, where the former encodes the feature map in each channel independently and the latter encodes the feature maps across all channels using a 1×1 kernel. Then, we use a 1×1 convolution operation with a sigmoid activation to generate the channel attention map A_c with the original input length. Compared to global convolution under the same inputs ($2688 \times 14 \times 14$), our method can reduce the number of parameters and the computational cost by more than 99.8%.

3.2.2. Global Spatial Attention Block

As a complement to the channel attention, the spatial attention aims to emphasize the spatial dimension of the input feature maps. It can learn the relationships among different spatial locations and generate a spatial weight map to find where feature pixels are important across all channels. The computation of a spatial weight map follows Equation (2) with a composite function and a sigmoid activation. In the previous works [27,28], the composite function is realized by a series of dilated convolutions or a large convolution with channel-wise pooling operations. The former may cause the gridding effect due to discontinuous feature encoding, while the latter has a limited receptive field, as an oversized convolution kernel easily leads to high model complexity.

In our work, we follow the global channel attention and use the depthwise and spatially separable convolutions to encode the input feature maps with a global receptive field. A key difference is that we add the spatial padding (P_w or P_h) along the horizontal or vertical direction to the feature maps before each convolution operation. The purpose is to preserve sufficient spatial cues by keeping the feature maps at a $1/2$ downsampling rate during encoding. Then, we use a $1 \times 2 \times 2$ deconvolution operation with a stride of two and a sigmoid activation to generate the spatial weight map A_s with the input spatial size.

3.3. Within-Task Feature Fusion Scheme

In the early RM-CNN model [24], the head architecture is typically designed as a stacked encoding structure, as shown in Figure 3a. Through 4 stacked 3×3 convolutions with 256 channels, the input features are encoded with increasing receptive field size to capture the spatial context. However, the number of stacked layers is limited using a single gradient flow between them. Too many stacked layers would cause the vanishing gradient problem during training. In addition, this stacked encoding scheme easily leads to excessive parameters and computational cost. Figure 3b shows a dense encoding scheme in DenseNet [29], which introduces dense skip connections into a predefined composite unit containing successive 1×1 and 3×3 convolutions with 128 and 64 channels, respectively. The dense encoding scheme can effectively improve the gradient flow of a network by adding $i + 1$ layer connections when performing the i-th encoding operation. However, the #Params and MACC of each encoding operation are increased by 4096 and $4096 \times W_r \times H_r$, respectively. This would result in a rapidly increasing model complexity if the encoding operation is performed multiple times.

In our work, we combine the advantages of the above two encoding schemes and design a new within-task encoding scheme based on deep feature fusion among layers. As shown in Figure 3c, we replace the composite unit of the dense encoding scheme with a new composite unit consisting of three consecutive 3×3 DS-Convs with progressively reduced channels of {192, 128, 64}. The output of each encoding operation is linked to the

outputs of previous DS-Convs operations using multiple skip connections. Compared to the previous encoding schemes, this design has the following three major strengths:

- Further improvement of gradient flow: In the proposed scheme, $3i + 1$ layer connections are added when the i-th encoding operation is performed. This feature can further improve the gradient flow of the downstream subnetwork during training.
- Lowerintroduction of model complexity: We replace standard convolutions with depthwise separable convolutions to reduce the overall model complexity of the proposed scheme. Compared to the stacked and dense encoding schemes under the same inputs ($2688 \times 14 \times 14$), our method can reduce the number of parameters and MACC by about 98.1% and 97.5%, respectively.
- Better capture of multi-scale context: In each encoding operation, the outputs contain the intermediate features with different receptive fields. Taking advantage of the lightweight design, we can perform the proposed encoding operation multiple times to capture more features with different scales of context information.

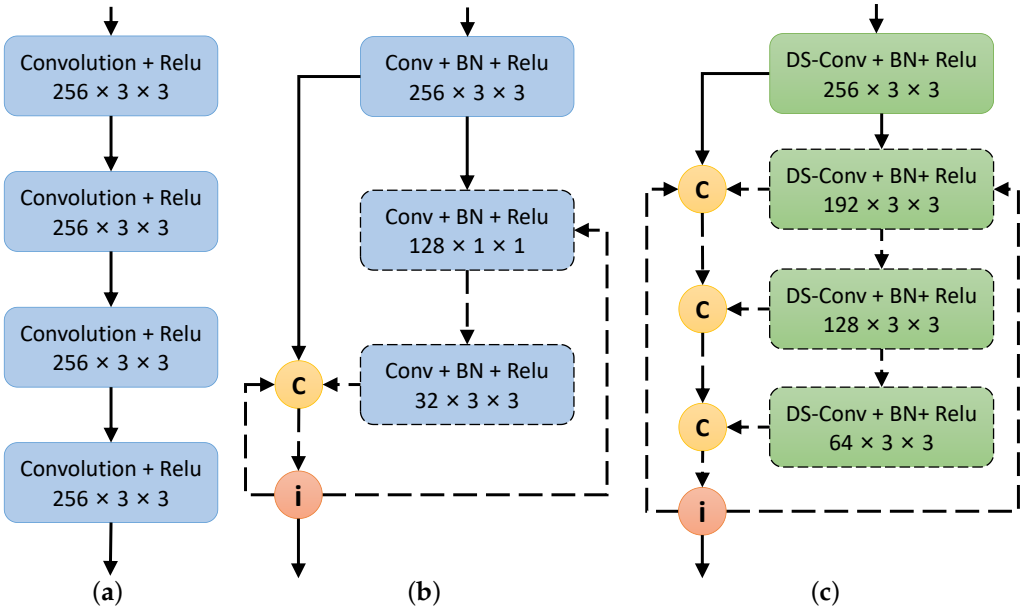

Figure 3. Overview of the stacked encoding scheme in Mask R-CNN (**a**), the dense encoding scheme in DenseNet (**b**), and the proposed feature-fused scheme (**c**). Note: "c" denotes a concatenation, and "i" denotes the output of the i-th encoding operation.

3.4. Location Reweighting Strategy

Instead of directly obtaining the points with the maximum probability values on the landmark score maps, we adopt a location reweighting strategy to account for the contribution of each predicted location with a probability value greater than 0.5. The strategy is formulated as follows:

$$\begin{bmatrix} L_x^n \\ L_y^n \end{bmatrix} = \frac{\sum_{ij} I(M_{ij}^n) M_{ij}^n \begin{bmatrix} i \\ j \end{bmatrix}}{\sum_{ij} I(M_{ij}^n) M_{ij}^n}, \tag{3}$$

where L_x^n and L_y^n denote the coordinates of the n-th predicted facial landmark on a detected region. M_{ij}^n is a probability value at the pixel location (i, j) of the corresponding score map. $I \in \{0, 1\}$ denotes an indicator function that is 0 if the input value is less than 0.5 and 1 otherwise.

3.5. Implementation Detail

Following the work [30], we train the region detection subnetwork using a softmax loss and a smooth L1 loss for the facial region classification and the bounding box regression, respectively. For the anchor setting, we use nine reference boxes with three scales $\{160, 208, 256\}$ and three aspect ratios $\{0.5, 1, 2\}$. In our model, we formulate the landmark localization task as a pixel-wise classification problem on the location response maps and train the landmark localization subnetwork using a multinomial cross-entropy loss with the per-pixel softmax operation. Each input image is resized to have a shortest side length of 256 pixels, and the training images are augmented using three scale ratios $\{0.5, 1, 2\}$ and three rotation angles $\{-30, 0, 30\}$. Each set of annotated facial landmarks is transformed into the ground-truth bounding box that tightly covers a facial region. The size of the region-wise features from the network backbone is uniformly set to 14×14 by the RoIAlign operation. The landmark localization subnetwork receives $2688 \times 14 \times 14$ region-wise features and outputs 120×120 landmark score maps. During training, we first fine-tune the region detection subnetwork based on the pre-trained ResNet-50 model and then train the landmark localization subnetwork from scratch. We use a stochastic gradient descent (SGD) optimizer with a mini-batch size of 2 and an initial learning rate of 0.001, which is decreased by a factor of 0.1 every 50k iterations. The momentum and weight decay are set to 0.9 and 5×10^{-4}, respectively. Our code will be made publicly available (https://github.com/MUST-AI-Lab/RDFN, accessed on 11 September 2023).

4. Experiments

4.1. Datasets and Settings

To demonstrate the effectiveness of the proposed model, we perform extensive experiments on three common datasets described as follows:

- 300W [4]: The 300W dataset consists of several popular datasets, such as the HELEN [31], LFPW [32] and AFW [33] datasets, and has been widely used to evaluate the face-alignment algorithms. It contains 3148 training images and 689 test images with 68 annotated landmarks. The test images are divided into the challenge subset (135 images) and the common subset (554 images).
- AFLW [5]: The AFLW dataset is an in-the-wild dataset containing 24,386 faces with large variations in head pose. Following the work [34], the AFLW dataset is split into the AFLW-Full and AFLW-Frontal datasets with 19 reduced landmarks. The AFLW-Full dataset contains 20,000 training images and 4384 test images, of which 1314 near-frontal test images are collected in the AFLW-Frontal dataset.
- COFW [2]: The COFW dataset provides 1852 unconstrained faces with different occlusions, which are divided into 1345 training images and 507 test images. Each face image is annotated with 29 landmarks and the corresponding occlusion state.

In our experiments, we follow most previous studies and use the normalized mean error (NME) with a specific normalized distance to evaluate the accuracy of the methods on different datasets. We also use the failure rate (FR) metric by setting a maximum NME of 10%. The number of parameters (#Params) and the number of floating point operations (FLOPs) are used to measure the model size and computational cost, respectively.

4.2. Comparison with Existing Methods on Common Datasets

In this section, we compare the proposed method (RDFN) with existing methods, including conventional CSR-based methods, such as SDM [1], RCPR [2] and ESR [13], DNN-based methods without reinitialization, such as TCDCN [18], HRNet [19] and AWing [7], and recent DNN-based methods with reinitialization, such as RCEN [11], PicassoNet [9], and SLPT [12]. For a fair comparison, we follow the train–test setting of common datasets to report all the methods. In the experiments, RDFN_{od} denotes the proposed method using the detected bounding boxes from the region detection subnetwork for region-wise feature extraction, while RDFN_{gt} is the one using the ground-truth bounding boxes in the landmark localization subnetwork.

4.2.1. Results on 300W

Following the 300W dataset setting [4], we normalize the NME metric using the interocular distance of a face. Table 1 reports the NMEs of the methods on the common subset, challenging subset and full set of 300W. We find that DNN-based methods have a clear performance advantage and dominate the state-of-the-art results compared to the conventional CSR-based method on different subsets. Our method achieves competitive performance compared to recent DNN-based methods with reinitialization and shows a promising potential of the RM-CNN architecture for facial landmark localization. We also find that the performance of RDFN_{od} is slightly weaker than that of RDFN_{gt} on all the subsets. This suggests that the detected bounding boxes of our method are suitable for the region-wise feature extraction in the landmark localization subnetwork.

Table 1. Comparison of NME (%) on the common subset, challenging subset and full set of 300W with 68 landmarks.

	Method	Common Subset	Challenging Subset	Full Set
Conventional CSR-Based Method	SDM [1]	5.57	15.40	7.50
	RCPR [2]	6.18	17.26	8.35
	ESR [13]	5.28	17.00	7.58
	ERT [35]	-	-	6.40
	LBF [6]	4.95	11.98	6.32
	CFSS [15]	4.73	9.98	5.76
DNN-Based Method w/o Reinitialization	MDM [16]	4.83	10.14	5.88
	TCDCN [18]	4.80	8.60	5.54
	RAR [17]	4.12	8.35	4.94
	SAN [36]	3.34	6.60	3.98
	LAB [37]	2.98	5.19	3.49
	ODN [38]	3.56	6.67	4.17
	HRNet [19]	2.87	5.15	3.32
	AWing [7]	2.72	4.52	3.07
	3FabRec [8]	3.36	5.74	3.82
	LGSA [39]	2.92	5.16	3.36
	SD-HRNet [21]	2.93	5.32	3.40
DNN-Based Method w/ Reinitialization	TSR [23]	4.36	7.42	4.96
	DAN [10]	3.19	5.24	3.59
	RCEN [11]	3.26	6.84	3.96
	PicassoNet [9]	3.03	5.81	3.58
	SLPT [12]	2.75	4.90	3.17
Ours	RDFN_{od}	2.82	5.19	3.28
	RDFN_{gt}	2.79	5.12	3.25

4.2.2. Results on AFLW

Due to the presence of different profile faces in AFLW, we follow the work [34] to use a face size as the normalized distance of the NME metric. Table 2 reports the NMEs of the methods on the test set of AFLW-Full and AFLW-Frontal. We find that our method has better performance than recent DNN-based methods with or without reinitialization. This means that a carefully designed RM-CNN architecture can achieve an impressive advantage when detecting a small number of face keypoints. Moreover, the experimental results show that the NMEs of RDFN_{od} have a small gap of no more than 0.06% when compared to those of RDFN_{gt} on both datasets.

Table 2. Comparison of NME (%) on the test set of AFLW-Full and AFLW-Frontal with 19 landmarks.

	Method	AFLW-Full	AFLW-Frontal
Conventional CSR-Based Method	SDM [1]	4.05	2.94
	RCPR [2]	3.73	2.87
	ERT [35]	4.35	2.75
	LBF [6]	4.25	2.74
	CFSS [15]	3.92	2.68
	CCL [34]	2.72	2.17
DNN-Based Method w/o Reinitialization	SAN [36]	1.91	1.85
	LAB [37]	1.85	1.62
	Wing [40]	1.65	-
	ODN [38]	1.63	1.38
	AWing [7]	1.53	1.38
	3FabRec [8]	1.84	1.59
DNN-Based Method w/Reinitialization	TSR [23]	2.17	-
	RCEN [11]	2.11	1.69
	PicassoNet [9]	1.59	1.30
Ours	$RDFN_{od}$	1.48	1.25
	$RDFN_{gt}$	1.42	1.21

4.2.3. Results on COFW

To validate the robustness of our method on occluded face images, we perform an evaluation on the test set of COFW and compare it with several popular conventional methods and recent DNN-based methods. Following the setting of previous works, we report the NME results normalized by the interocular distance as well as the corresponding FR results in Table 3. Our method outperforms most of the classical DNN-based methods, such as LAB and HRNet, and achieves competitive performance when compared to recent state-of-the-art methods, such as LGSA and SLPT. Similar to the observations on 300W and AFLW, our method generates the expected region bounding boxes with a performance close to that using the ground-truth bounding boxes. In Figure 4, we show the example results of our method on 300W, AFLW and COFW.

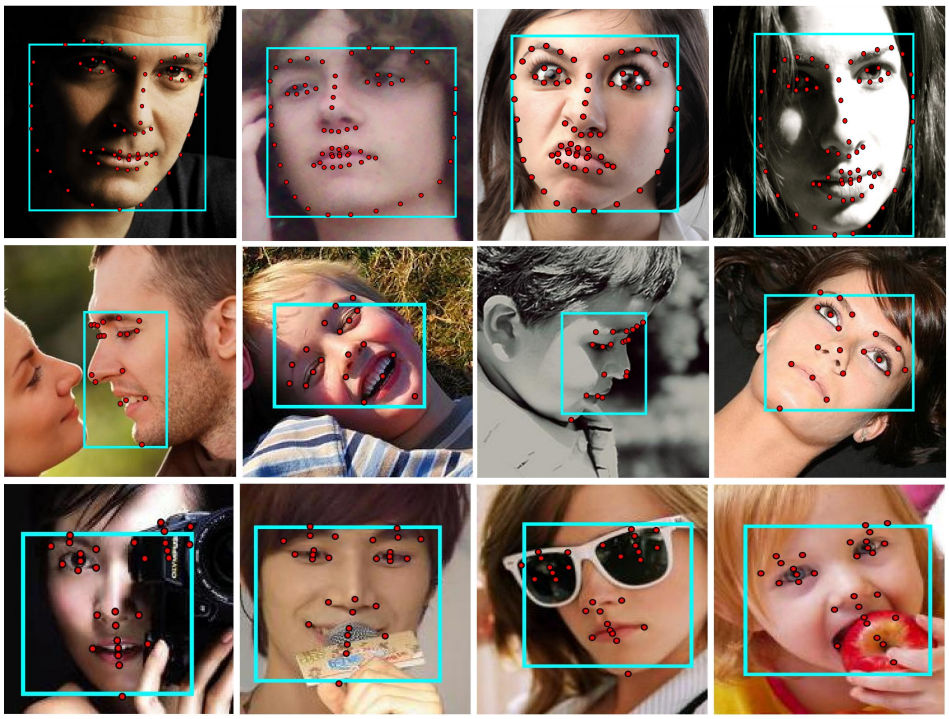

Figure 4. Example results of our facial landmark localization method. Top row: the results on 300W (68 landmarks). Second row: the results on AFLW (19 landmarks). Bottom row: the results on COFW (29 landmarks). The blue boxes indicate the detected region bounding boxes.

Table 3. Comparison of NME (%) and FR (%) on the test set of COFW with 29 landmarks.

	Method	NME (%)	FR (%)
Conventional CSR-Based Method	SDM [1]	11.14	-
	RCPR [2]	8.50	20.00
	ESR [13]	11.20	-
DNN-Based Method w/o Reinitialization	TCDCN [18]	8.05	-
	RAR [17]	6.03	4.14
	ECT [41]	5.98	4.54
	LAB [37]	3.92	0.39
	HRNet [19]	3.45	0.19
	LGSA [39]	3.13	0.002
	SD-HRNet [21]	3.61	0.12
DNN-Based Method w/Reinitialization	DRDA [22]	6.46	6.00
	RCEN [11]	4.44	2.56
	SLPT [12]	3.32	0.00
Ours	$RDFN_{od}$	3.43	0.18
	$RDFN_{gt}$	3.36	0.10

4.2.4. Complexity Analysis

In Table 4, we report the types of network backbone, #Params, and FLOPs of our method and recent DNN-based methods. We find that most of the previous methods use large backbone models, such as ResNet-50/152 and Hourglass, and achieve state-of-the-art performance at that time with a large number of parameters and high computational cost. In recent years, some works have improved the model complexity using efficient backbone models, such as HRNet, or network architecture search (NAS) techniques, such as SD-HRNet and PicassoNet. In our work, we use a reduced ResNet-50 network as the backbone module and introduce efficient feature fusion schemes. Our model is more lightweight than recent state-of-the-art methods, such as SLPT while achieving competitive performance on several common datasets. However, our model still has much room for improvement in terms of model complexity when compared to recent small models using the NAS method.

Table 4. Comparison of types of network backbone, #Params, and FLOPs in different methods.

	Method	Backbone	#Params (M)	FLOPs (G)
DNN-Based Method w/o Reinitialization	SAN [36]	ResNet-152	57.4	10.7
	LAB [37]	Hourglass	25.1	19.1
	Wing [40]	ResNet-50	25	-
	HRNet [19]	HRNetV2-W18	9.3	4.3
	AWing [7]	Hourglass	24.15	26.79
	LGSA [39]	Hourglass	18.64	-
	SD-HRNet [21]	-	0.98	0.59
DNN-Based Method w/Reinitialization	PicassoNet [9]	-	1.96	0.11
	SLPT [12]	HRNetV2-W18	13.18	5.17
Ours	RDFN	ResNet-50-C4	10.66	4.38

4.3. Ablation Experiments

To evaluate the contribution of each component in the proposed RDFN, we perform the ablation experiments on the common subset, challenging subset and full set of 300W. We use a vanilla RM-CNN with only the stacked encoding scheme as the baseline model and compare it with other variants using the proposed components, including the global channel attention block (GCA), the global spatial attention block (GSA), the within-task feature fusion scheme (WFF), and the location reweighting strategy (LR).

As shown in Figure 5, we find that the performance of a vanilla RM-CNN is poor and worse than recent DNN-based methods. By introducing the GCA block with cross-task feature fusion from multiple layers, the RM-CNN model achieves a significant improvement with NME reduced by 1.26% and FR reduced by 5.7% on the 300W full set. As other components are added, the NMEs of the RM-CNN model are progressively improved,

which indicates the effectiveness of each component. Furthermore, we find that our method performs better on the simple and near-frontal face images and achieves lower NME and FR results on the common subset, while there is still considerable room for improvement on the complex face images from the challenging subset.

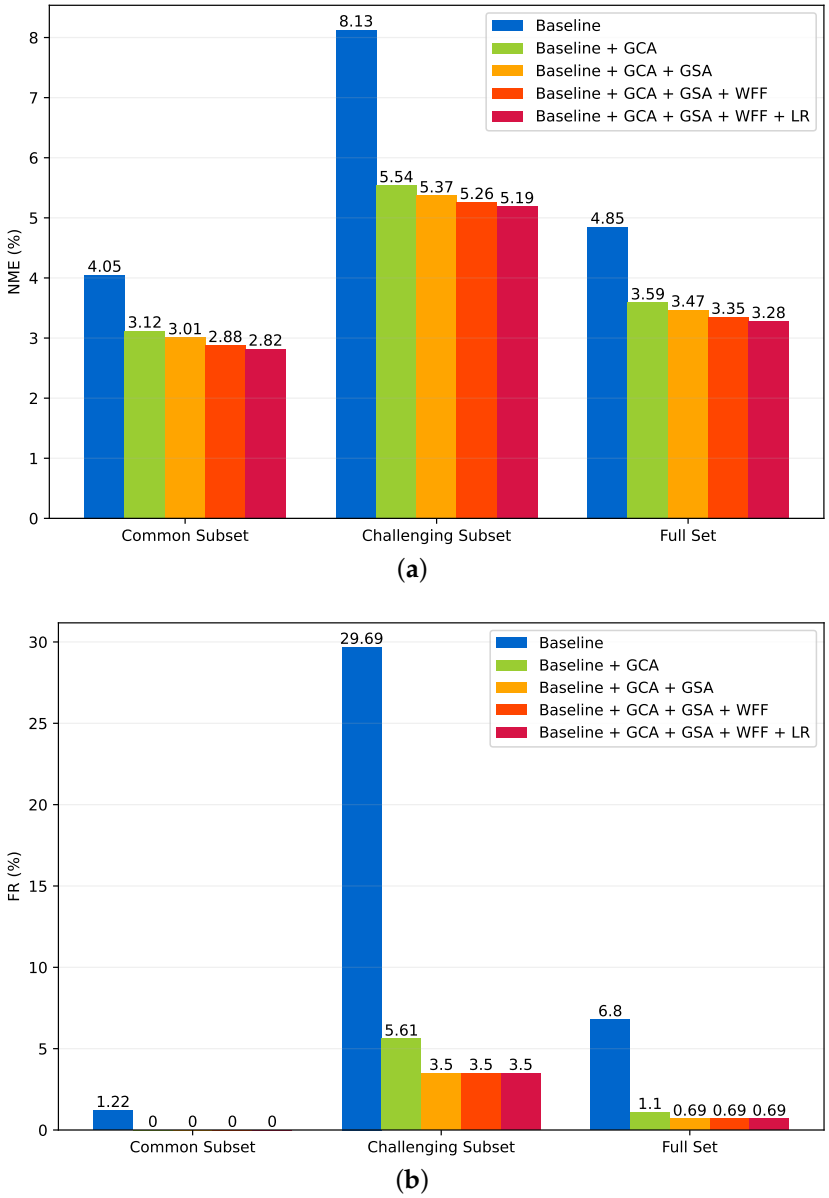

Figure 5. Comparison of NME (%) (**a**) and FR (%) (**b**) on the 300W common subset, challenging subset and full set. Note: "Baseline" denotes a vanilla RM-CNN model, "GCA" denotes the global channel attention block, "GSA" denotes the global spatial attention block, "WFF" denotes the within-task feature fusion scheme, "LR" denotes the location reweighting strategy.

5. Conclusions

As a crucial step in face applications, facial landmark localization has been widely studied to achieve increasingly accurate performance. However, how to make the landmark detector adaptive to different unconstrained inputs is still a challenging problem. In this paper, we propose an end-to-end region-aware network architecture that aims to simultaneously solve the input initialization problem and the facial landmark localization task. To balance the model complexity and inference performance, we design two lightweight feature fusion schemes to enhance the discriminative ability of the feature representation.

Furthermore, we consider the overall contribution of landmark score maps and present a location reweighting strategy to transform the score maps into 2D landmark coordinates in the post-processing stage. Extensive experiments demonstrate the effectiveness of the proposed model and related components. Our method achieves competitive performance compared to recent state-of-the-art methods while avoiding expensive computational costs. Nevertheless, our model still has potential room for improvement in terms of the model complexity when used in real-time applications. In future work, we will investigate the design of more efficient and effective backbone and feature fusion modules to further improve the accuracy and efficiency of our model.

Author Contributions: Conceptualization, X.L.; methodology, X.L.; software, X.L.; validation, X.L. and Y.L.; formal analysis, X.L. and Y.L.; investigation, X.L.; data curation, X.L.; writing—original draft preparation, X.L.; writing—review and editing, X.L. and Y.L.; visualization, X.L.; project administration, Y.L.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the China Postdoctoral Science Foundation under Grant 2020M683157, in part by Science and Technology Development Fund of Macau (0004/2020/A1, 0070/2020/AMJ), and in part by Guangdong Provincial Key R&D Programme: 2019B010148001.

Data Availability Statement: Not applicable.

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

References

1. Xiong, X.; De la Torre, F. Supervised descent method and its applications to face alignment. In Proceedings of the CVPR, Portland, OR, USA, 23–28 June 2013; pp. 532–539.
2. Burgos-Artizzu, X.P.; Perona, P.; Dollár, P. Robust face landmark estimation under occlusion. In Proceedings of the ICCV, Sydney, NSW, Australia, 1–8 December 2013; pp. 1513–1520.
3. Yan, J.; Lei, Z.; Yi, D.; Li, S. Learn to combine multiple hypotheses for accurate face alignment. In Proceedings of the ICCVW, Washington, DC, USA, 2–8 December 2013; pp. 392–396.
4. Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; Pantic, M. 300 Faces in-the-wild challenge: The first facial landmark localization challenge. In Proceedings of the ICCVW, Washington, DC, USA, 2–8 December 2013; pp. 397–403.
5. Koestinger, M.; Wohlhart, P.; Roth, P.M.; Bischof, H. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 2144–2151.
6. Ren, S.; Cao, X.; Wei, Y.; Sun, J. Face alignment at 3000 fps via regressing local binary features. In Proceedings of the CVPR, Columbus, OH, USA, 23–28 June 2014; pp. 1685–1692.
7. 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, Republic of Korea, 27 October–2 November 2019; pp. 6971–6981.
8. 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, 13–19 June 2020; pp. 6110–6120.
9. Wen, T.; Ding, Z.; Yao, Y.; Wang, Y.; Qian, X. Picassonet: Searching adaptive architecture for efficient facial landmark localization. *IEEE Trans. Neural Netw. Learn. Syst.* **2022** , 1–12. [CrossRef]
10. Kowalski, M.; Naruniec, J.; Trzcinski, T. Deep alignment network: A convolutional neural network for robust face alignment. In Proceedings of the CVPR, Faces-in-the-Wild Workshop/Challenge, Honolulu, HI, USA, 21–26 July 2017; Volume 3, p. 6.
11. Lin, X.; Liang, Y.; Wan, J.; Lin, C.; Li, S.Z. Region-based Context Enhanced Network for Robust Multiple Face Alignment. *IEEE Trans. Multimed.* **2019**, *21*, 3053–3067. [CrossRef]
12. Xia, J.; Qu, W.; Huang, W.; Zhang, J.; Wang, X.; Xu, M. Sparse Local Patch Transformer for Robust Face Alignment and Landmarks Inherent Relation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4052–4061.
13. Cao, X.; Wei, Y.; Wen, F.; Sun, J. Face alignment by explicit shape regression. *Int. J. Comput. Vis.* **2014**, *107*, 177–190. [CrossRef]
14. Feng, Z.H.; Hu, G.; Kittler, J.; Christmas, W.; Wu, X.J. Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting. *IEEE Trans. Image Process.* **2015**, *24*, 3425–3440. [CrossRef]
15. Zhu, S.; Li, C.; Loy, C.; Tang, X. Face alignment by coarse-to-fine shape searching. In Proceedings of the CVPR, Boston, MA, USA, 7–12 June 2015; pp. 4998–5006.
16. 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 CVPR, Las Vegas, NV, USA, 27–30 June 2016; pp. 4177–4187.

17. 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 ECCV, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 57–72.
18. Zhang, Z.; Luo, P.; Loy, C.C.; Tang, X. Learning deep representation for face alignment with auxiliary attributes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 918–930. [CrossRef]
19. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-resolution representations for labeling pixels and regions. *arXiv* **2019**, arXiv:1904.04514.
20. 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; pp. 483–499.
21. Lin, X.; Zheng, H.; Zhao, P.; Liang, Y. SD-HRNet: Slimming and Distilling High-Resolution Network for Efficient Face Alignment. *Sensors* **2023**, *23*, 1532. [CrossRef] [PubMed]
22. Zhang, J.; Kan, M.; Shan, S.; Chen, X. Occlusion-free face alignment: Deep regression networks coupled with de-corrupt autoencoders. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 3428–3437.
23. Lv, J.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 CVPR, Honolulu, HI, USA, 21–26 July 2017; Volume 1, p. 4.
24. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 770–778.
26. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
27. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
28. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. Bam: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514.
29. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
30. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
31. Le, V.; Brandt, J.; Lin, Z.; Bourdev, L.; Huang, T.S. Interactive facial feature localization. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 679–692.
32. Belhumeur, P.N.; Jacobs, D.W.; Kriegman, D.J.; Kumar, N. Localizing parts of faces using a consensus of exemplars. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2930–2940. [CrossRef] [PubMed]
33. Zhu, X.; Ramanan, D. Face detection, pose estimation, and landmark localization in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2879–2886.
34. Zhu, S.; Li, C.; Loy, C.C.; Tang, X. Unconstrained face alignment via cascaded compositional learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 3409–3417.
35. Kazemi, V.; Sullivan, J. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the CVPR, Columbus, OH, USA, 23–28 June 2014; pp. 1867–1874.
36. Dong, X.; Yan, Y.; Ouyang, W.; Yang, Y. Style aggregated network for facial landmark detection. In Proceedings of the CVPR, Salt Lake City, UT, USA, 18–22 June 2018; Volume 2, p. 6.
37. 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 CVPR, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2129–2138.
38. 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.
39. Gao, P.; Lu, K.; Xue, J.; Shao, L.; Lyu, J. A coarse-to-fine facial landmark detection method based on self-attention mechanism. *IEEE Trans. Multimed.* **2021**, *23*, 926–938. [CrossRef]
40. Feng, Z.H.; Kittler, J.; Awais, M.; Huber, P.; Wu, X.J. Wing loss for robust facial landmark localisation with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2235–2245.
41. Zhang, H.; Li, Q.; Sun, Z.; Liu, Y. Combining data-driven and model-driven methods for robust facial landmark detection. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 2409–2422. [CrossRef]