sensors

MDPI

Article

SD-HRNet: Slimming and Distilling High-Resolution Network for Efficient Face Alignment

Xuxin Lin [1,2,†], Haowen Zheng [2,†], Penghui Zhao [2] and Yanyan Liang [2,*]

[1] Zhuhai Da Heng Qin Technology Development Co., Ltd., Zhuhai 519000, China
[2] Faculty of Innovation Engineering, Macau University of Science and Technology, Macau 999078, China
* Correspondence: yyliang@must.edu.mo
† These authors contributed equally to this work.

Abstract: Face alignment is widely used in high-level face analysis applications, such as human activity recognition and human–computer interaction. However, most existing models involve a large number of parameters and are computationally inefficient in practical applications. In this paper, we aim to build a lightweight facial landmark detector by proposing a network-level architecture-slimming method. Concretely, we introduce a selective feature fusion mechanism to quantify and prune redundant transformation and aggregation operations in a high-resolution supernetwork. Moreover, we develop a triple knowledge distillation scheme to further refine a slimmed network, where two peer student networks could learn the implicit landmark distributions from each other while absorbing the knowledge from a teacher network. Extensive experiments on challenging benchmarks, including 300W, COFW, and WFLW, demonstrate that our approach achieves competitive performance with a better trade-off between the number of parameters (0.98 M–1.32 M) and the number of floating-point operations (0.59 G–0.6 G) when compared to recent state-of-the-art methods.

Keywords: face alignment; knowledge distillation; network pruning; lightweight model

Citation: Lin, X.; Zheng, H.; Zhao, P.; Liang, Y. SD-HRNet: Slimming and Distilling High-Resolution Network for Efficient Face Alignment. *Sensors* **2023**, *23*, 1532. https://doi.org/10.3390/s23031532

Academic Editors: Maurizio Caon and Hai-Ning Liang

Received: 27 December 2022
Revised: 20 January 2023
Accepted: 22 January 2023
Published: 30 January 2023

1. Introduction

Face alignment, also known as facial landmark detection, aims at locating a set of semantic points on a given face image. It usually serves as a critical step in many face applications, such as face recognition [1], expression analysis [2], and driver-status tracking [3], which are significant components of human–computer interaction systems. As an example, face alignment is used to generate a canonical face in the preprocessing of face recognition [4,5]. In the past decade, there have been many methods and common datasets reported in the literature [6–22] to promote the development of face alignment. Nevertheless, it remains a challenging task to develop an efficient and robust facial landmark detector that performs well in various unconstrained scenarios.

In early works, the methods [6–9] based on cascaded regression made significant progress on face alignment. They could learn a mapping function to iteratively refine the estimated landmark positions from an initial face shape. Despite the success of the methods for near-frontal face alignment, their performance was dramatically degraded on challenging benchmarks. The main reason is that the methods use handcrafted features and simply learned regression methods, which are weak to take full advantage of data for the accurate shape mapping on unconstrained faces.

With the development of deep learning on computer vision, convolutional neural network (CNN)-based methods have achieved impressive performance for unconstrained face alignment. Most existing works focus on improving the accuracy of the landmark localization by utilizing large backbone networks (e.g., VGG-16 [23], ResNet-50/152 [24], and Hourglass [25]). Although the networks have powerful feature extraction ability, they involve many parameters and a high computational cost and are difficult to apply in resource-limited environments.

Recently, some researchers have tended to balance the accuracy and efficiency of a facial landmark detector. They either train a small model from scratch [26,27] or use knowledge distillation (KD) for model compression [28–31]. The former aims to design a lightweight network combined with an effective learning strategy, while the latter considers how to apply the KD technique to transfer the dark knowledge from a large network to a small one. However, the methods are not flexible enough to adapt to different computing resources as they usually rely on a fixed and carefully designed network structure.

Inspired by the works [32,33] of neural architecture search and neural network pruning for image classification, in which a compact target network was derived from a large supernetwork, we attempted to search for a lightweight face alignment network from a dynamically learned neural architecture. Concretely, we first trained a high-resolution supernetwork based on the structure of HRNet [34]. In this network, a lightweight selective feature fusion (LSFF) block was designed to quantify the importance of the built-in transformation and aggregation operations. Then, we optionally pruned the redundant operations or even the entire blocks to obtain a slimmed network. To reduce the performance gap between the slimmed network and the supernetwork, we developed a triple knowledge distillation scheme, where two peer student networks with masked inputs could learn the ensemble of landmark distributions while receiving the knowledge from a frozen teacher network. In this paper, our main contributions are summarized as follows:

- We propose a flexible network-level architecture slimming method that can quantify and reduce the redundancy of the network structure to obtain a lightweight facial landmark detector adapted to different computing resources.
- We design a triple knowledge distillation scheme, in which a slimmed network could be improved without additional complexity by jointly learning the implicit landmark distribution from a teacher network and two peer student networks.
- Extensive experimental results on challenging benchmarks demonstrate that our approach achieves a better trade-off between accuracy and efficiency than recent state-of-the-art methods (see Figure 1).

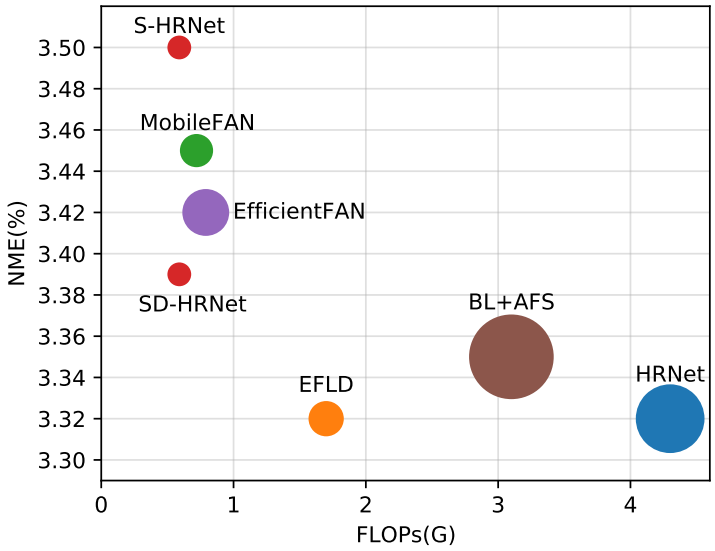


Figure 1. Comparison of the computational cost (i.e., FLOPs) and the performance (i.e., NME) on 300W between the proposed approach and existing state-of-the-art methods. The size of a circle represents the number of parameters. Our approach (SD-HRNet) achieves a better trade-off between accuracy and efficiency than its counterparts.

The rest of this paper is organized as follows: Section 2 provides a review of related works about existing face alignment methods. In Section 3, we describe the detail of our

proposed slimming and distillation methods. Section 4 shows the experimental results and analysis on common datasets. Finally, we give a brief conclusion in Section 5.

2. Related Work

In this section, we provide a detailed review of the related methods on face alignment.

2.1. Conventional Face Alignment

In the early literature [6–9], the cascaded regression method was popular and widely used to predict facial landmark positions by resolving a regression problem. The representative methods included SDM [6], ESR [7], LBF [8], and CFSS [9]. The main differences among the methods were the choices of extracted features and the landmark regression methods. SDM used the scale-invariant feature transform (SIFT) as a feature descriptor applied to a cascaded linear regression model. ESR was a two-stage boosted regression method to predict the landmark coordinates by using the shape-indexed features. LBF combined the random forest algorithm with local binary features to accelerate the landmark localization process. To avoid a local optimum due to poor initialization, CFSS exploited hybrid image features to estimate the landmark positions in a coarse-to-fine manner. These methods were weak to detect landmarks on unconstrained face images due to the use of handcrafted features and simply learned regression methods. In our work, we build a CNN model to jointly learn the deep feature extraction and facial landmark heatmap regression.

2.2. Large CNN-Based Face Alignment

In recent years, there have been some advanced approaches reported in the literature [10–19], which have exploited large CNN models to drastically improve the landmark localization accuracy. Wu and Yang [10] proposed a deep variation leveraging network (DVLN), which contained two strongly coupled VGG-16 networks for landmark prediction and candidate decision. Lin et al. [11,12] adopted a classic two-stage detection architecture [35] based on the VGG-16 backbone for joint face detection and alignment. Feng et al. [13] and Dong et al. [14] applied the ResNet-50 [24] and ResNet-152 [24] networks, respectively, as the feature extraction module in the landmark detection process. The stacked hourglass network [25] is a popular CNN backbone used in recent state-of-the-art works [15,16,18] to generate features with multiscale information. Xia et al. [19] combined the HRNet backbone with a transformer structure to achieve a coarse-to-fine face alignment framework. The methods had high accuracy on challenging benchmarks, but inevitably required a large number of parameters and a high computational cost. Our approach only utilizes the large CNN model (HRNet) as a teacher network and adopts a lightweight model for face alignment.

2.3. Lightweight CNN-Based Face Alignment

Due to the limited application of large CNN models, some researchers have begun to study the lightweight network design for face alignment. Bulat et al. [26] applied the network quantization technique to construct a binary hourglass network. Guo et al. [27] trained a lightweight network consisting of the MobileNetV2 [36] blocks by using an auxiliary 3D pose estimator. To utilize the learning ability of large models, some recent works [28–31] used the teacher-guided KD technique to make a small student network learn the dark knowledge from a large teacher network. The student networks were usually based on the existing lightweight networks (e.g., MobileNetV2, EfficientNet-B0 [37], and HRNetV2-W9 [34]), while the teacher networks use the large CNN models (e.g., ResNet-50, EfficientNet-B7 [37], and HRNetV2-W18 [34]) as the network backbone. It is worth mentioning that the KD technique was also applied to improve a facial landmark detector [38–40] by mining the spatial–temporal relation from unlabeled video data. Inspired by the student-guided KD [41] that made student networks learn from each other without a teacher network, we introduce a student-guided learning strategy into the original KD framework, which can generate more robust supervision knowledge for learning landmark

distribution. Moreover, our student network is derived from a supernetwork and thus has a more flexible structure than other handcrafted models.

3. Methods

As illustrated in Figure 2, our approach is a two-stage process consisting of a network-level architecture slimming and triple knowledge distillation, which results in a lightweight facial landmark detector.

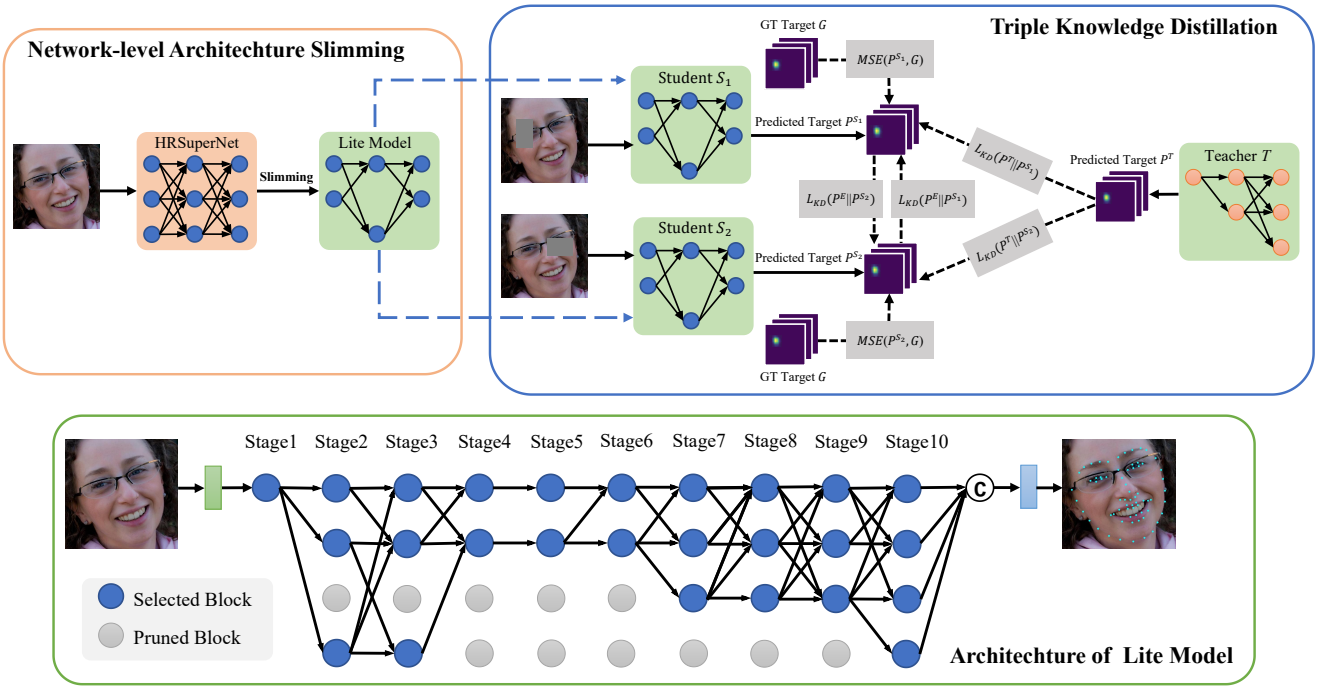


Figure 2. Illustration of the proposed slimming and distillation procedures for face alignment. A lightweight model is first obtained by slimming the HRSuperNet. Then, the lightweight model is refined in a triple knowledge distillation scheme consisting of two peer student networks and a teacher network. We visualize the architecture of the lightweight model trained on the 300W dataset, where the redundant TA operations and LSFF blocks are pruned.

3.1. Network-Level Architecture Slimming

Our high-resolution supernetwork (HRSuperNet) follows a similar structure to HRNet in Figure 3 and begins from a stem that is composed of two 3×3 convolutions with a stride of 2. The spatial resolution is downsampled to $H/4 \times W/4$, where H and W denote the height and width of an input image $I \in \mathbb{R}^{3 \times H \times W}$. The main body consists of ten stages maintaining the high-resolution representations throughout the network. Different from HRNet, the supernetwork contains a single-resolution LSFF block with a downsampling ratio of 1 in the first stage and repeats four-resolution blocks with downsampling ratios of $\{1, 1/2, 1/4, 1/8\}$ from the beginning of the second stage. Each block has four stacked mobile inverted bottleneck convolutions (MBConvs [36]) with a 3×3 kernel size and an expansion ratio of 1. The design could make the supernetwork keep a larger architecture space but fewer parameters and lower computational cost than HRNet. Except for the first stage, the LSFF block is designed to transform and aggregate features from the previous stage and generate new features as inputs to the next stage. The process is formulated as follows:

$$\mathbf{Y}_{i>1,k} = E\Big(\sum_{j=1}^{J_{i-1}} \alpha_{i,j}^{k} T(\mathbf{X}_{i-1,j}^{k})\Big), \tag{1}$$

where $\mathbf{Y}_{i,k}$ is the output of the kth block in the ith stage and $\mathbf{X}_{i-1,j}^{k}$ denotes the kth output from the jth block in the $(i-1)$th stage. J_{i-1} is the number of blocks in the $(i-1)$th stage.

T represents a transformation operation that is either a 1×1 convolution with a stride of 1 and a bilinear interpolation for upsampling, a sequence of 3×3 convolutions with a stride of 2 for downsampling, or an identity shortcut connection. E denotes a feature encoding operation implemented by the stacked MBconvs. The factor α is used as the weight of each transformation operation to participate in the follow-up aggregation process. The head in the supernetwork consists of two 1×1 convolutions with a stride of 1 and generates the landmark heatmaps $\mathbf{P} \in \mathbb{R}^{N \times M \times H/4 \times W/4}$ when receiving N samples with M facial landmark points.

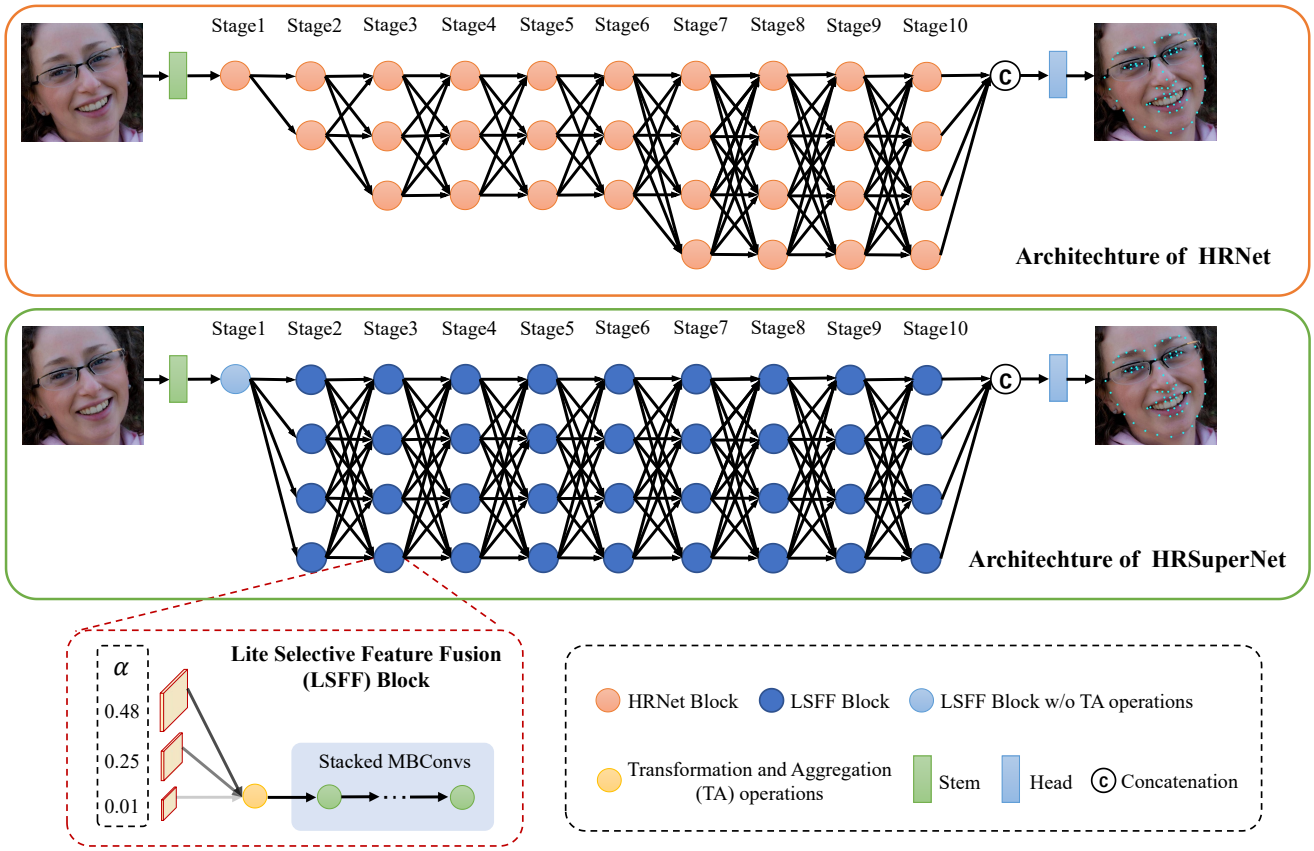


Figure 3. Detailed structures of HRNet and HRSuperNet. The proposed lightweight selective feature fusion (LSFF) block is composed of the transformation and aggregation (TA) operations with different importance factors α and stacked mobile inverted bottleneck convolutions (MBConvs).

During the training, we make the supernetwork learn the landmark heatmap regression along with the subnetwork architecture search by imposing an L1 regularization on α to enforce the sparsity of the operations with few contributions to the network. Formally, the overall training loss is:

$$L = \sum_{n=1}^{N} \sum_{m=1}^{M} \frac{MSE(\mathbf{P}_{n,m}, \mathbf{G}_{n,m})}{N \times M} + \lambda \sum_{i=2}^{I} \sum_{j=1}^{J_{i-1}} \sum_{k=1}^{J_i} |\alpha_{i,j}^{k}|, \tag{2}$$

where $MSE(\mathbf{P}_{n,m}, \mathbf{G}_{n,m})$ denotes the standard mean square error between the predicted heatmap $\mathbf{P}_{n,m}$ and the ground-truth heatmap $\mathbf{G}_{n,m}$ of the mth landmark in the nth sample. The ground-truth heatmap is generated by applying a 2D Gaussian centered on the ground-truth location of each landmark. λ is the weight to balance the MSE and the L1 penalty term. I and J_i denote the number of stages and the number of blocks in the ith stage, respectively. We first train the supernetwork by alternately optimizing the importance factors and the network weights until they converge. Then, we prune the redundant transformation and

aggregation operations in the LSFF blocks, where the corresponding factors are smaller than a given pruning threshold. Note that the entire block is discarded if all the associated operations are pruned.

3.2. Triple Knowledge Distillation

In our distillation scheme, we adopt the slimmed network as the peer student networks S_1 and S_2 and use the pretrained HRNet as the teacher network T. To increase the model diversity, we use the occluded images with a random-sized mask as the inputs of the student networks.

Specially, we define a KD loss for a network to learn the landmark distribution from another network as follows:

$$L_{KD}(\mathbf{P}^2||\mathbf{P}^1) = \sum_{n=1}^{N}\sum_{m=1}^{M}\frac{D_{KL}(S(\mathbf{P}^2_{n,m})||S(\mathbf{P}^1_{n,m}))}{N\times M}, \tag{3}$$

where D_{KL} is the Kullback–Leibler (KL) divergence to measure the distance of the landmark distributions from $S(\mathbf{P}^1)$ to $S(\mathbf{P}^2)$, and S is the softmax function working on the predicted landmark heatmaps $\mathbf{P}^1$ and $\mathbf{P}^2$.

During the training, we use MSE and L_{KD} as the main criterion to make the student networks learn the explicit landmark distribution from the ground-truth heatmap, while allowing them to learn the implicit landmark distribution from their ensemble predictions and the output of the teacher network. The overall training loss of a student network S_i is formulated as:

$$\begin{aligned}\mathbf{P}^E &= (\mathbf{P}^{S_1}+\mathbf{P}^{S_2})/2,\\ L_{S_i} &= \sum_{n=1}^{N}\sum_{m=1}^{M}\frac{MSE(\mathbf{P}^{S_i}_{n,m},\mathbf{G}_{n,m})}{N\times M}\\ &\quad+\lambda_1 L_{KD}(\mathbf{P}^E||\mathbf{P}^{S_i})+\lambda_2 L_{KD}(\mathbf{P}^T||\mathbf{P}^{S_i}),\end{aligned} \tag{4}$$

where $\mathbf{P}^{S_1}$, $\mathbf{P}^{S_2}$, and $\mathbf{P}^T$ denote the predicted landmark heatmaps of S_1, S_2, and T, respectively. The weights λ_1 and λ_2 are used to balance MSE and the KD losses.

4. Experiments

4.1. Datasets

We conducted experiments on three challenging datasets including 300W [20], COFW [21], and WFLW [15].

300W: It consists of the HELEN [42], LFPW [43], AFW [44], XM2VTS [45], and IBUG [20] datasets, where each face has 68 landmarks. The training set contains 3148 images and the test set has 689 images divided into the challenge subset (135 images) and the common subset (554 images). Masked 300W [46] is a supplement to the 300W dataset for testing. This dataset mainly includes masked faces with over 50% of occlusion.

COFW: It contains 1852 face images with different degrees of occlusion including 1345 training images and 507 test images. Each face image has 29 annotated landmarks.

WFLW: There are 7500 images for training and 2500 images for testing where the test set includes six subsets: large pose (326 images), illumination (698 images), occlusion (736 images), blur (773 images), make-up (206 images), and expression (314 images).

4.2. Evaluation Metrics

We followed previous works and used the normalized mean error (NME) to evaluate the performance of the facial landmark detection:

$$NME = \sum_{n=1}^{N}\sum_{m=1}^{M}\frac{\|\mathbf{p}_{n,m}-\mathbf{g}_{n,m}\|_2}{N\times M\times d} \tag{5}$$

where $\mathbf{p}_{n,m}$ and $\mathbf{g}_{n,m}$ denote the coordinate vectors of the predicted landmark and the ground-truth landmark, respectively. d is the interocular distance. We also report the failure rate by setting a maximum NME of 10%. The number of parameters (#Params) and the number of floating-point operations (FLOPs) were used to measure model size and computational cost, respectively.

4.3. Implementation Detail

Following the work [15], all the faces were cropped based on the provided bounding boxes and resized to 256×256. We augmented the data by a 1.0 ± 0.25 scaling, ± 30-degree rotation, and random flipping with a probability of 50%. The pseudocode in Algorithm 1 shows the training pipeline of our approach in the slimming and distilling stages.

Algorithm 1: SD-HRNet Algorithm

Input: The training set D_T, initialized importance factor α and network weight w, training epochs N, pruning threshold p, pretrained teacher network T
Output: Two lightweight networks S_1 and S_2

for $i = 1$ *to* N **do**
  for *Mini-batch* D_t *in* D_T **do**
    Calculate the loss L by Eqaution (2)
    Update α by gradient descent:
    $\alpha = \alpha - \nabla_\alpha L$
  end
  for *Mini-batch* D_t *in* D_T **do**
    Calculate the loss L by Equation (2)
    Update w by gradient descent:
    $w = w - \nabla_w L$
  end
end
Obtain lightweight networks S_1 and S_2 by p
Initialize importance factors in S_1 and S_2:
$\alpha_1, \alpha_2 = \alpha$
Initialize network weights in S_1 and S_2:
$w_1, w_2 = w$
for $i = 1$ *to* N **do**
  for *Minibatch* D_t *in* D_T **do**
    Calculate the losses L_{S_1} and L_{S_2} by Equation (4)
    Update w_1 and w_2 by gradient descent:
    $w_1 = w_1 - \nabla_{w_1} L_{S_1}$
    $w_2 = w_2 - \nabla_{w_2} L_{S_2}$
  end
end

Slimming stage: We alternatively optimized the importance factors and network weights for 60 epochs. To optimize the importance factors, we used the Adam optimizer with the learning rates of 1.8×10^{-4} on 300W and WFLW, and 3.5×10^{-4} on COFW. The weight λ was set to 5×10^{-5}. To update the network weights, we used the Adam optimizer with a momentum of 0.9 and weight decay of 4×10^{-5}. The initial learning rate was 1×10^{-4} on 300W and COFW, and 2×10^{-4} on WFLW, which was dropped by a factor of 0.1 in 40 and 50 epochs. The pruning threshold was set to 0.0017 on 300W, 0.0042 on COFW, and 0.002 on WFLW. The batch size was set to 16 on 300W and COFW, and 32 on WFLW.

Distilling stage: We jointly fine-tuned two slimmed networks for 60 epochs. The settings of the optimizer, learning rate and batch size were the same as those in the slimming

stage. The weights λ_1 and λ_2 were set to 4 and 1 on 300W, 3 and 0.1 on COFW, and 0.3 and 0.1 on WFLW.

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

In this section, we compare our approach with the recent state-of-the-art methods on 300W, COFW, and WFLW. S-HRNet is the slimmed network from HRSuperNet. SD-HRNet$_1$ and SD-HRNet$_2$ denote two refined S-HRNets through the proposed distillation scheme. We report the average results with the standard deviation of SD-HRNet$_1$ and SD-HRNet$_2$ from training them five times over different seeds.

4.4.1. Results on 300W

We report the #Params and FLOPs in Table 1 as well as the NME on the 300W subsets in Table 2. Compared to the advanced models (e.g., LAB and SLPT) with large backbones, our method (SD-HRNet) had far fewer parameters and FLOPs while achieving competitive or even better performance. Compared to HRNet, SD-HRNet only increased the NME by about 2% but reduced the #Params by 89.5% and the FLOPs by 86.3%. We showed that SD-HRNet achieved fewer parameters (0.98 M parameters) and a lower computational cost (0.59 G FLOPs) than existing lightweight models. Moreover, we proved the effectiveness of the proposed slimming and distillation approaches as the #Params and FLOPs of HRSuperNet were reduced by 70.0% and 52.8%, respectively, and the NME of S-HRNet was reduced by about 3%. Table 3 shows the performance of the methods on Masked 300W. We found that SD-HRNet had an obvious improvement for occluded faces due to the introduction of masked inputs. Although our method obtained a competitive NME compared to most previous methods, it underperformed the recent state-of-the-art methods [47,48] focusing on the occlusion problem.

Table 1. Comparison of different methods in backbone, #Params, and FLOPs.

Method	Backbone	#Params (M)	FLOPs (G)
DVLN [10]	VGG-16 [23]	132.0	14.4
Wing+PDB [13]	ResNet-50 [24]	25	3.8
SAN [14]	ResNet-152 [24]	57.4	10.7
LAB [15]	Hourglass [25]	25.1	19.1
HRNet [34]	HRNetV2-W18 [34]	9.3	4.3
AWing [16]	Hourglass [25]	24.15	26.79
BL+AFS [17]	-	14.29	3.10
LGSA [18]	Hourglass [25]	18.64	15.69
SLPT [19]	HRNetV2-W18 [34]	13.18	5.17
SRN [47]	Hourglass [25]	19.89	-
GlomFace [48]	-	-	13.48
MobileFAN [28]	MobileNetV2 [36]	2.02	0.72
EfficientFAN [29]	EfficientNet-B0 [37]	4.19	0.79
EFLD [30]	HRNetV2-W9 [34]	2.3	1.7
mnv2$_{KD}$ [31]	MobileNetV2 [36]	2.4	0.6
HRSuperNet	HRSuperNet	3.27	1.25
SD-HRNet (300W)	S-HRNet	0.98	0.59
SD-HRNet (COFW)	S-HRNet	1.05	0.60
SD-HRNet (WFLW)	S-HRNet	1.32	0.60

Table 2. Comparison of NME (%) on 300W: common subset, challenge subset, and full set.

Method	Common	Challenge	Full
ODN [49]	3.56	6.67	4.17
SAN [14]	3.34	6.60	3.98
LAB [15]	2.98	5.19	3.49
HRNet [34]	2.87	5.15	3.32
AWing [16]	2.72	4.52	3.07
BL+AFS [17]	2.89	5.23	3.35
LGSA [18]	2.92	5.16	3.36
SAAT [46]	2.87	5.03	3.29
SRN [47]	3.08	5.86	3.62
GlomFace [48]	2.79	4.87	3.20
SLPT [19]	2.75	4.90	3.17
MobileFAN [28]	2.98	5.34	3.45
EfficientFAN [29]	2.98	5.21	3.42
EFLD [30]	2.88	5.03	3.32
$mnv2_{KD}$ [31]	3.56	6.13	4.06
HRSuperNet	3.00	5.28	3.45
S-HRNet	3.02	5.44	3.50
$SD\text{-}HRNet_1$	2.93 ± 0.01	5.32 ± 0.05	3.40 ± 0.01
$SD\text{-}HRNet_2$	2.94 ± 0.01	5.33 ± 0.02	3.41 ± 0.01

Table 3. Comparison of NME (%) on Masked 300W: common subset, challenge subset, and full set.

Method	Common	Challenge	Full
CFSS [9]	11.73	19.98	13.35
DHGN [50]	8.98	12.19	9.61
SBR [38]	8.72	13.28	9.6
SHG [25]	8.17	13.52	9.22
MDM [51]	7.66	11.67	8.44
FHR [52]	7.02	11.28	7.85
LAB [15]	6.07	9.59	6.76
SAAT [46]	5.42	11.36	6.58
SRN [47]	5.78	9.28	6.46
GlomFace [48]	5.29	8.81	5.98
HRSuperNet	14.03	20.52	15.30
S-HRNet	19.18	29.37	21.17
$SD\text{-}HRNet_1$	6.43 ± 0.25	11.05 ± 0.51	7.34 ± 0.28
$SD\text{-}HRNet_2$	6.23 ± 0.23	10.72 ± 0.25	7.11 ± 0.20

4.4.2. Results on COFW

Table 4 shows the NME and the failure rate for a maximum NME of 10% on the COFW test set. Our method performed better than some classic works (e.g., RAR and DAC-CSR) for partially occluded face alignment and achieved a competitive accuracy against recent state-of-the-art methods (e.g., LGSA and SLPT). Compared to HRNet, the NME of SD-HRNet was slightly increased by about 5% while the #Params and FLOPs were reduced by 88.7% and 86.0%, respectively. Moreover, SD-HRNet was still more lightweight than recent small models and had similar landmark detection performance.

Table 4. Comparison of NME (%) and failure rate (%) for a maximum NME of 10% on the COFW test set.

Method	NME (%)	Failure Rate (%)
HPM [53]	7.50	13.00
CCR [54]	7.03	10.9
DRDA [55]	6.46	6.00
RAR [56]	6.03	4.14
DAC-CSR [57]	6.03	4.73
Wing+PDB [13]	5.07	3.16
LAB [15]	3.92	0.39
HRNet [34]	3.45	0.19
LGSA [18]	3.13	0.002
SLPT [19]	3.32	0.00
MobileFAN [28]	3.66	0.59
EfficientFAN [29]	3.40	0.00
EFLD [30]	3.50	0.00
$mnv2_{KD}$ [31]	4.11	2.36
HRSuperNet	3.74	0.59
S-HRNet	3.69	0.20
$SD\text{-}HRNet_1$	3.61 ± 0.02	0.12 ± 0.16
$SD\text{-}HRNet_2$	3.63 ± 0.03	0.20 ± 0.17

4.4.3. Results on WFLW

In Table 5, we report the NME on the WFLW test set and six subsets. Our method significantly outperformed conventional cascaded regression methods (e.g., SDM and CFSS) and some classic large models (e.g., LAB and Wing+PDB). However, we found that there was a slightly bigger performance gap between SD-HRNet and recent large models than the results on 300W and COFW. The reason might be that learning the dense landmark regression relies on a large network capacity. Compared to the advanced lightweight models, SD-HRNet achieved the third-best NME in most cases with a better trade-off of model size (1.32 M parameters) and computational cost (0.6 G FLOPs).

Table 5. Comparison of NME (%) on the WFLW test set and 6 subsets: pose, expression, illumination, make-up, occlusion, and blur.

Method	Test	Pose	Expression	Illumination	Make-Up	Occlusion	Blur
ESR [58]	11.13	25.88	11.47	10.49	11.05	13.75	12.20
SDM [6]	10.29	24.10	11.45	9.32	9.38	13.03	11.28
CFSS [9]	9.07	21.36	10.09	8.30	8.74	11.76	9.96
DVLN [10]	6.08	11.54	6.78	5.73	5.98	7.33	6.88
LAB [15]	5.27	10.24	5.51	5.23	5.15	6.79	6.32
Wing+PDB [13]	5.11	8.75	5.36	4.93	5.41	6.37	5.81
HRNet [34]	4.60	7.94	4.85	4.55	4.29	5.44	5.42
AWing [16]	4.36	7.38	4.58	4.32	4.27	5.19	4.96
LUVLi [59]	4.37	7.56	4.77	4.30	4.33	5.29	4.94
LGSA [18]	4.28	7.63	4.33	4.16	4.27	5.33	4.95
$mnv2_{KD}$ [31]	8.57	15.06	8.81	8.15	8.75	9.92	9.40
MobileFAN [28]	4.93	8.72	5.27	4.93	4.70	5.94	5.73
EFLD [30]	4.74	8.41	5.01	4.71	4.57	5.70	5.45
EfficientFAN [29]	4.54	8.20	4.87	4.39	4.54	5.42	5.04
HRSuperNet	4.83	8.45	5.10	4.80	4.85	5.79	5.53
S-HRNet	4.98	8.68	5.33	4.86	4.88	5.87	5.70
$SD\text{-}HRNet_1$	4.93 ± 0.01	8.63 ± 0.03	5.31 ± 0.05	4.81 ± 0.03	4.76 ± 0.02	5.73 ± 0.02	5.56 ± 0.03
$SD\text{-}HRNet_2$	4.96 ± 0.03	8.66 ± 0.10	5.35 ± 0.04	4.82 ± 0.05	4.81 ± 0.04	5.76 ± 0.04	5.61 ± 0.06

In Figure 4, we show the number of frames per second (FPS) of our method using different batch sizes on 300W. Due to the very small resource consumption, SD-HRNet could process more than 500 samples per second. In Figure 5, we give some example results on the common datasets and show the accurate landmark localization of our method on various unconstrained faces. All the experiments were implemented with PyTorch on a single TITAN Xp GPU.

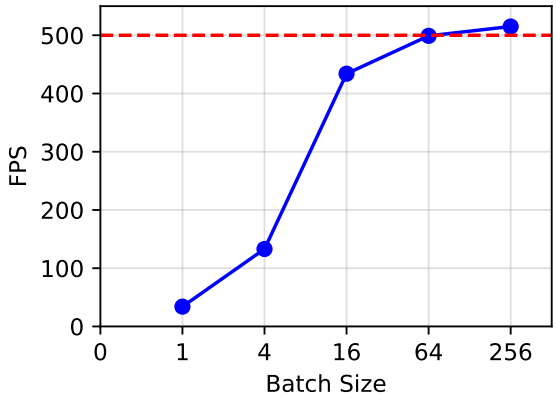


Figure 4. FPS of our method using different batch sizes on 300W.

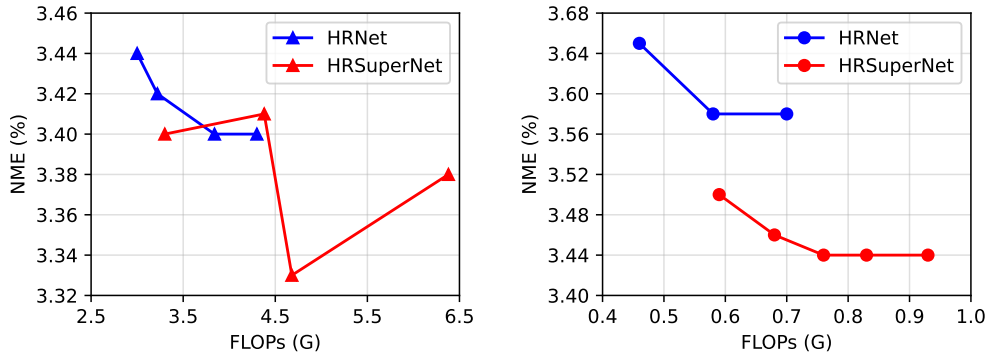
Figure 5. Example results of our method for face alignment. Top row: results on 300W (68 points). Second row: results on WFLW (98 points). Bottom row: results on COFW (29 points).

4.5. Ablation Study

In this section, we conduct an ablation study on 300W and analyze the effect of the proposed components.

4.5.1. HRNet vs. HRSuperNet

To verify the rationality of our supernetwork, we trained HRNet and HRSuperNet on 300W without pretraining and used them as the supernetwork to generate a series of slimmed networks. The original residual units [34] or stacked MBConvs were used as the feature encoding operation in the proposed LSFF block. As seen from Figure 6, most networks derived from HRSuperNet had a lower NME than HRNet when their FLOPs were similar, which suggested that a larger architecture space was more likely to generate better subnetworks.



(**a**) (**b**)

Figure 6. Comparison of HRNet and HRSuperNet based on original residual units (**a**) or stacked MBConvs (**b**), which were used as the supernetwork in the proposed slimming method. We obtained a series of slimmed networks with different NME and FLOPs on the 300W full set by using different pruning thresholds.

4.5.2. KD Components

In Table 6, we show the effect of different KD components in our distillation scheme for the performance on 300W. We observed that each component incrementally led to the improvement of the slimmed network. It suggested that the combination of the teacher-guided KD and the student-guided KD was an effective way for the implicit knowledge transfer. In addition, the introduction of masked inputs could increase the diversity of student networks and make them learn robust landmark distribution from each other.

Table 6. NME (%) of our method using different KD components on the 300W full set.

Teacher	Peer Student	Masked Inputs	NME (%)
×	×	×	3.50
✓	×	×	3.46
✓	✓	×	3.44
✓	✓	✓	3.39

4.6. Visualization of the Architectures

We visualize the slimmed architecture trained on 300W in Figure 2 and the other two architectures on COFW and WFLW in Figure 7. The proposed selective feature fusion mechanism could result in different network structures from a unified architecture space, which were adapted to different datasets and landmark detection tasks. For example, the architectures from 300W and COFW tended to preserve more high-resolution blocks from the first and second branches than the architecture from WFLW. In addition, we found that more than 94% of the blocks in HRSuperNet were utilized by the slimmed architectures. It suggested that the designed architecture space was reasonable to cover most cases for generating an efficient face alignment network.

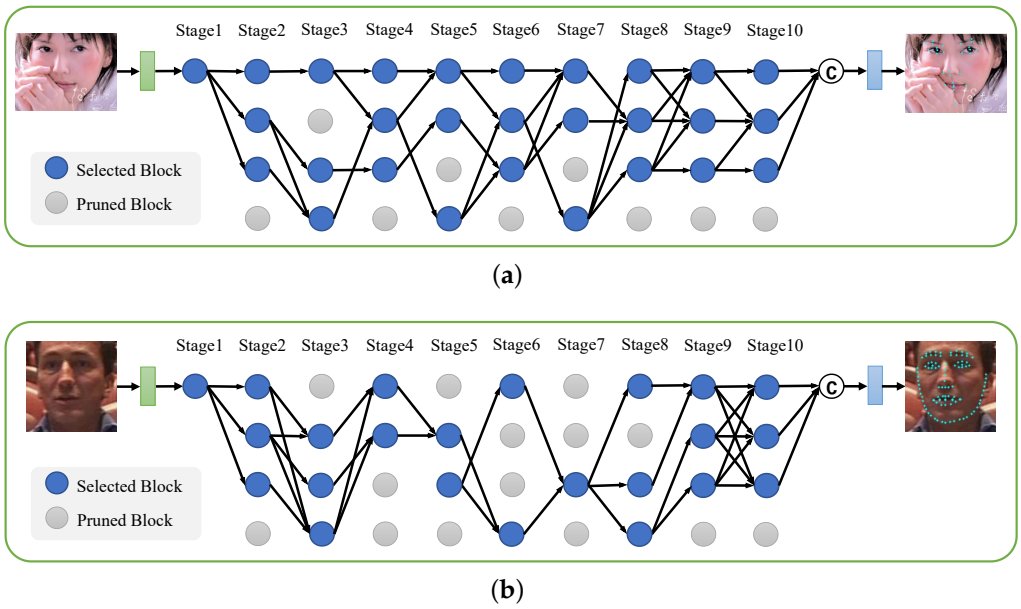


Figure 7. Visualization of the slimmed architectures trained on the COFW (**a**) and WFLW (**b**) datasets.

5. Conclusions

In this paper, we proposed a network-level slimming method and a hybrid knowledge distillation scheme, which could work together to generate an efficient and accurate facial landmark detector. Compared to existing handcrafted models, our model achieved competitive performance with a better trade-off between model size (0.98 M–1.32 M parameters) and computational cost (0.59 G–0.6 G FLOPs). In addition, our method was more flexible in practical application through an adaptive architecture search technique, which could be applied to real-time human–computer interaction systems under different resource-limited environments. Nevertheless, there was still a performance gap between our method and recent state-of-the-art large models, especially for the dense or strongly occluded landmark detection task. In future work, we will explore how to design a more reasonable architecture search space to improve the upper bound of performance and extend our method to other computer vision tasks such as human pose estimation and semantic segmentation.

Author Contributions: Conceptualization, X.L.; methodology, X.L. and H.Z.; software, H.Z.; validation, H.Z.; formal analysis, Y.L.; investigation, P.Z.; data curation, P.Z.; writing—original draft preparation, X.L. and H.Z.; writing—review and editing, X.L. and Y.L.; visualization, P.Z.; 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 (0010/2019/AFJ, 0025/2019/AKP, 0004/2020/A1, and 0070/2020/AMJ), and in part by Guangdong Provincial Key R&D Programme: 2019B010148001.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code is available at https://github.com/MUST-AI-Lab/SD-HRNet (accessed on 27 December 2022).

Acknowledgments: The authors are grateful for the discussion with Hongqiang Wei. He also provided project supervision and key technical support.

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

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional neural network
HRSuperNet	High-resolution super network
S-HRNet	Slimmed network from HRSuperNet
SD-HRNet	Refined S-HRNet using triple knowledge distillation
LSFF	Lightweight selective feature fusion
TA	Transformation and aggregation
MBConvs	Mobile inverted bottleneck convolutions
KD	Knowledge distillation
KL	Kullback–Leibler
2D	Two-dimensional
FLOPs	Number of floating-point operations
NME	Normalized mean error
#Params	Number of parameters
FPS	Frames per second
SIFT	Scale-invariant feature transform
M	Mega
G	Giga

References

1. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014, pp. 1701–1708.
2. Pantic, M.; Rothkrantz, L.J.M. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1424–1445. [CrossRef]
3. Jabbar, R.; Shinoy, M.; Kharbeche, M.; Al-Khalifa, K.; Krichen, M.; Barkaoui, K. Driver drowsiness detection model using convolutional neural networks techniques for android application. In Proceedings of the 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), Doha, Qatar, 2–5 February 2020; pp. 237–242.
4. Taskiran, M.; Kahraman, N.; Erdem, C.E. Face recognition: Past, present and future (a review). *Digital Signal Process.* **2020**, *106*, 102809. [CrossRef]
5. Adjabi, I.; Ouahabi, A.; Benzaoui, A.; Taleb-Ahmed, A. Past, present, and future of face recognition: A review. *Electronics* **2020**, *9*, 1188. [CrossRef]
6. Xiong, X.; la Torre, F. Supervised descent method and its applications to face alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 532–539.
7. Cao, X.; Wei, Y.; Wen, F.; Sun, J. Face alignment by explicit shape regression. *Int. J. Comput. Vision* **2014**, *107*, 177–190. [CrossRef]
8. Ren, S.; Cao, X.; Wei, Y.; Sun, J. Face alignment at 3000 fps via regressing local binary features. In Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014; pp. 1685–1692.
9. Zhu, S.; Li, C.; Loy, C.; Tang, X. Face alignment by coarse-to-fine shape searching. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4998–5006.
10. Wu, W.; Yang, S. Leveraging intra and inter-dataset variations for robust face alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2096–2105.
11. Lin, X.; Wan, J.; Xie, Y.; Zhang, S.; Lin, C.; Liang, Y.; Guo, G.; Li, S. Task-Oriented Feature-Fused Network With Multivariate Dataset for Joint Face Analysis. *IEEE Trans. Cybern.* **2019**, *50*, 1292–1305. [CrossRef] [PubMed]
12. 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]
13. 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake, UT, USA, 18–23 June 2018; pp. 2235–2245.
14. Dong, X.; Yan, Y.; Ouyang, W.; Yang, Y. Style aggregated network for facial landmark detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake, UT, USA, 18–23 June 2018; pp. 379–388.
15. 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake, UT, USA, 18–23 June 2018; pp. 2129–2138.
16. 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 (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6971–6981.
17. Wang, T.; Tong, X.; Cai, W. Attention-based face alignment: A solution to speed/accuracy trade-off. *Neurocomputing* **2020**, *400*, 86–96. [CrossRef]

18. 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]
19. 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 (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4052–4061.
20. 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, NSW, Australia, 2–8 December 2013; pp. 397–403.
21. 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, NSW, Australia, 1–8 December 2013; pp. 1513–1520.
22. Ghiasi, G.; Fowlkes, C.C. Occlusion coherence: Detecting and localizing occluded faces. *arXiv* **2015**, arXiv:1506.08347.
23. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. In Proceedings of the British Machine Vision Conference 2014, Lenton, Nottingham, UK, 1–5 September 2014.
24. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. 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.
26. Bulat, A.; Tzimiropoulos, G. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In Proceedings of IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3706–3714.
27. Guo, X.; Li, S.; Yu, J.; Zhang, J.; Ma, J.; Ma, L.; Liu, W.; Ling, H. PFLD: A practical facial landmark detector. *arXiv* **2019**, arXiv:1902.10859v2.
28. Zhao, Y.; Liu, Y.; Shen, C.; Gao, Y.; Xiong, S. Mobilefan: Transferring deep hidden representation for face alignment. *Pattern Recognit.* **2020**, *100*, 107114. [CrossRef]
29. Gao, P.; Lu, K.; Xue, J.; Lyu, J.; Shao, L. A facial landmark detection method based on deep knowledge transfer. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**. [CrossRef] [PubMed]
30. Sha, Y. Efficient Facial Landmark Detector by Knowledge Distillation. In Proceedings of 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), Jodhpur, India, 15–18 December 2021; pp. 1–8.
31. Fard, A.P.; Mahoor, M.H. Facial landmark points detection using knowledge distillation-based neural networks. *Comput. Vis. Image Underst.* **2022**, *215*, 103316. [CrossRef]
32. Liu, H.; Simonyan, K.; Yang, Y. Darts: Differentiable architecture search. *arXiv* **2018**, arXiv:1806.09055.
33. Yu, J.; Yang, L.; Xu, N.; Yang, J.; Huang, T. Slimmable Neural Networks. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
34. 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.
35. 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.
36. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
37. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning (PMLR), Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
38. Dong, X.; Yu, S.; Weng, X.; Wei, S.; Yang, Y.; Sheikh, Y. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; p. 360–368.
39. Zhu, C.; Liu, H.; Yu, Z.; Sun, X. Towards omni-supervised face alignment for large scale unlabeled videos. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 13090–13097.
40. Zhu, C.; Li, X.; Li, J.; Dai, S.; Tong, W. Multi-sourced Knowledge Integration for Robust Self-Supervised Facial Landmark Tracking. *IEEE Trans. Multimed.* **2022**. [CrossRef]
41. Zhang, Y.; Xiang, T.; Hospedales, T.M.; Lu, H. Deep mutual learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4320–4328.
42. Le, V.; Brandt, J.; Lin, Z.; Bourdev, L.; Huang, T.S. Interactive facial feature localization. In Proceedings of the 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 679–692.
43. 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]
44. 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.

45. Messer, K.; Matas, J.; Kittler, J.; Luettin, J.; Maitre, G.; et al. XM2VTSDB: The extended M2VTS database. In Proceedings of the International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA), Washington, DC, USA, 22–23 March 1999; Volume 964, pp. 965–966.
46. 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, QC, Canada, 10–17 October 2021; pp. 11751–11760.
47. Zhu, C.; Li, X.; Li, J.; Dai, S.; Tong, W. Reasoning structural relation for occlusion-robust facial landmark localization. *Pattern Recognit.* **2022**, *122*, 108325. [CrossRef]
48. Zhu, C.; Wan, X.; Xie, S.; Li, X.; Gu, Y. Occlusion-Robust Face Alignment Using a Viewpoint-Invariant Hierarchical Network Architecture. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022, pp. 11112–11121.
49. 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, 15–20 June 2019; pp. 3486–3496.
50. 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]
51. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4177–4187.
52. 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; pp. 8893–8900.
53. Ghiasi, G.; Fowlkes, C. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2385–2392.
54. Feng, Z.; Hu, G.; Kittler, J.; Christmas, W.; Wu, X. 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] [PubMed]
55. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3428–3437.
56. 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 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 57–72.
57. Feng, Z.; Kittler, J.; Christmas, W.; Huber, P.; Wu, X. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3681–3690.
58. Cao, X.; Wei, Y.; Wen, F.; Sun, J. Face alignment by explicit shape regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2887–2894.
59. 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 (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8236–8246.