entropy

MDPI

Article

Attention-Guided Huber Loss for Head Pose Estimation Based on Improved Capsule Network

Runhao Zhong [1], Li He [1,*], Hongwei Wang [1], Liang Yuan [1,2], Kexin Li [1] and Zhening Liu [1]

[1] School of Mechanical Engineering, Xinjiang University, Urumqi 830046, China; 107552101383@stu.xju.edu.cn (R.Z.); wanghongwei@stu.xju.edu.cn (H.W.); yuanliang@mail.buct.edu.cn (L.Y.); li_kexin@stu.xju.edu.cn (K.L.); 107552103899@stu.xju.edu.cn (Z.L.)
[2] School of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China
* Correspondence: xju_heli@xju.edu.cn

Abstract: Head pose estimation is an important technology for analyzing human behavior and has been widely researched and applied in areas such as human–computer interaction and fatigue detection. However, traditional head pose estimation networks suffer from the problem of easily losing spatial structure information, particularly in complex scenarios where occlusions and multiple object detections are common, resulting in low accuracy. To address the above issues, we propose a head pose estimation model based on the residual network and capsule network. Firstly, a deep residual network is used to extract features from three stages, capturing spatial structure information at different levels, and a global attention block is employed to enhance the spatial weight of feature extraction. To effectively avoid the loss of spatial structure information, the features are encoded and transmitted to the output using an improved capsule network, which is enhanced in its generalization ability through self-attention routing mechanisms. To enhance the robustness of the model, we optimize Huber loss, which is first used in head pose estimation. Finally, experiments are conducted on three popular public datasets, 300W-LP, AFLW2000, and BIWI. The results demonstrate that the proposed method achieves state-of-the-art results, particularly in scenarios with occlusions.

Keywords: head pose estimation; global attention block; self-attention routing; capsule network

Citation: Zhong, R.; He, L.; Wang, H.; Yuan, L.; Li, K.; Liu, Z. Attention-Guided Huber Loss for Head Pose Estimation Based on Improved Capsule Network. *Entropy* **2023**, *25*, 1024. https://doi.org/10.3390/e25071024

Academic Editor: Wei Li

Received: 4 May 2023
Revised: 29 June 2023
Accepted: 4 July 2023
Published: 5 July 2023

1. Introduction

The rapid development of artificial intelligence and the increasing prevalence of intelligent living have led to an increasing demand for robots to serve people in their daily lives. As an important component of service robot co-navigation, head pose estimation can provide information on human attention direction and intent, assisting robots in analyzing human behavior, and enabling them to possess the capability of social-aware navigation [1].

Head pose estimation refers to inferring the orientation of the head from a given image. It can be represented by a 3D vector that includes the pitch, roll, and yaw angles [2]. In recent years, extensive research on head pose estimation has driven the development of 3D reconstruction and interaction behavior analysis [3]. Head pose estimation has been widely used in many applications including virtual reality [4], fatigue driving detection [5], motion capture [6], and many other areas. In the driving system, it can determine the driver's attention and consciousness based on position information [7]. Head pose is associated with visual attention, allowing semantic cues to be combined with dialogue and human interaction to facilitate non-verbal communication in special places, as well as being an important cue for predicting the direction of pedestrian movement. Therefore, it is used in many computer vision systems such as gaze discrimination [8], augmented reality [9], human–computer interaction [10], and surveillance security [11]. Chen [12] proposed a system for analyzing human behavior and human interaction in meetings and workplaces using head posture estimation.

Head pose estimation from an image requires learning a mapping of two-dimensional space and three-dimensional space. Traditional head pose estimation methods use face localization and image cropping to reduce the influence of image-independent backgrounds on the detection target, mainly by using face alignment to similarly match the target image to the image sample [13]. However, head pose estimation in complex environments is a challenging problem. To obtain an effective representation of facial features in partially occluded scenarios, Wu [14] extracted pyramid HoG features from non-occluded facial sub-regions to estimate head pose. Wang [15] improved the performance of the model by synthesizing head pose images and augmenting the sample images under different lighting and occlusion conditions. Xing [16] introduced occluded regions into facial appearance, recovered facial shape from partially occluded facial appearances, and modeled various types of partial facial occlusions. These methods are typically limited to individual head pose estimation and have low computational efficiency.

Head pose estimation utilizes a single RGB image since it can be regarded as a classification problem. Most of the methods proposed in the past five years for head pose estimation are based on machine learning and convolutional/deep neural networks [17]. Compared with other deep neural networks, the main contribution of Convolutional Neural Networks (CNN) is their ability to effectively replicate features from all positions in the input image's spatial dimensions and use the learned features at other positions to achieve spatial reduction and local shared connectivity [18]. Nevertheless, CNNs often lack local equivariant features, resulting in weak generalization ability and loss of important object localization information, requiring additional parameters to construct deep networks. Hinton [19] proposed a novel form of cooperation between neurons using a new type of unit called capsules. In this unit, individual activations no longer represent the presence of specific features, but instead represent different properties of the same entity. Compared to traditional individual neurons, capsules consist of multiple neurons. On the convolution layer, the normal single-layer convolution is two-dimensional, while a single-layer capsule is three-dimensional. Each neuron within a capsule represents a specific attribute such as size, orientation, hue, texture, and other features. Consequently, capsules can capture more detailed information. As the head undergoes angular changes in the image, the neurons within the capsules also exhibit corresponding variations. Subsequently, a dynamic routing algorithm is employed to select and propagate capsules, thereby influencing the final output. Building on this idea, Sabor [20] proposed a capsule network that introduces capsules into the traditional CNN architecture. Each capsule represents a group of neurons that parameterize the instantiation parameters associated with different targets. Yang [21] proposed a head pose estimation method that employs capsule networks to filter candidate features and obtain representative features. However, the current capsule networks suffer from efficiency issues and inadequate representational capacity. They require many parameters, which inevitably obscure the intrinsic generalization ability that capsules should provide, leading to suboptimal accuracy in head pose estimation.

In this paper, we argue that head pose estimation methods suffer from low estimation accuracy and poor occlusion detection in complex environments. Therefore, we consider the limitations of the loss of spatial structure information and the inadequate generalization ability of capsule networks. To overcome these limitations, we propose a head pose estimation method based on an improved capsule network and attention-guided regression loss to enhance the accuracy of head pose estimation.

Specifically, the contributions of our work are as follows:

(1) We propose a detection model that combines CNN and capsule network for head pose estimation. The model includes feature extraction, feature mapping, and feature aggregation modules. A multi-level output structure backbone network is used to extract spatial structure and semantic information at different levels for improved estimation accuracy.

(2) In order to obtain more effective and representative features, we apply an improved feature extraction module to enhance the spatial weight of feature extraction, which

can obtain more spatial information and significantly improve the feature extraction performance and network efficiency for capturing the main features.

(3) To address the problem of loss of spatial structure information, we utilize an improved capsule network with reduced number of capsules and parameters to enhance the network's generalization ability. Moreover, we optimize the regression loss to improve the model's robustness.

(4) To validate the effectiveness of the proposed method, we conduct tests and ablation experiments on the AFLW2000 and BIWI datasets. The results demonstrated that the performance of our method outperformed previous methods significantly.

2. Related Work

Head pose estimation, as a critical technology in computer vision, has attracted extensive attention and research. Various methods for head pose estimation have been proposed up to now based on different types of data including 2D color images, 3D images, and depth images. Methods based on 3D images and depth images require special equipment, such as depth cameras or stereo cameras, which are expensive and computationally complex, making real-time applications difficult [22]. Therefore, usually, only RGB images are utilized for estimating and analyzing head pose. In addition, previous methods using depth cameras are only accurate at close distances [23]. Depending on whether facial landmarks detection is required, head pose estimation methods are divided into landmark-based methods and landmark-freed methods.

2.1. Landmark-Based Methods

Facial landmark detection is used to establish the mapping relationship between 3D space and 2D images for estimating head pose [24]. Dlib [25] employed a collection of regression trees to locate facial landmarks with real-time prediction speed. 3DDFA [26] converted the head into a dense 3D model and uses a CNN to fit the 3D model to an RGB image. This method can effectively deal with occlusion problems. Nikolaidis [27] proposed a novel method for detecting facial landmarks using a combination of Adaptive Hough Transform (AHT) [28] and template matching techniques. The detected landmarks are then utilized to calculate the horizontal head pose, based on the deformation of an equilateral triangle formed by the landmarks of the two eyes and mouth. This method has shown promising results in accurately estimating the horizontal head pose in various scenarios. To improve the accuracy of head pose estimation, Narayan [29] proposed a universal geometric model for horizontal head pose estimation and validated its effectiveness on multiple standard datasets. The geometric-based approach has a simple process and low time complexity, requiring only a few facial features to obtain suitable head pose estimates. FAN [30], an advanced landmark detection method, obtains multiscale information by merging block features across layers and is robust to occlusion and head pose. KEPLER [31] uses a modified google net architecture to simultaneously predict facial key-points and poses. The coarse pose supervision methods are used to improve landmark detection.

The accurate estimation of head pose based on facial landmarks and model matching requires sufficient accuracy in both facial detection and feature point labeling. However, due to a high complexity, computational cost, and low efficiency of the model, it is difficult to accurately estimate head pose using classification training methods. In practical applications, the accuracy of facial feature point detection can be significantly reduced by various interfering factors such as lighting changes, complex backgrounds, head rotation, and occlusion [32], which can even make it impossible to detect facial feature points. Therefore, the model-based head pose estimation method is not entirely accurate.

2.2. Landmark-Freed Methods

The landmark-freed methods train on different samples of poses to obtain a vector description of the target pose sample, which represents a mapping between the pose and its features. By converting the head pose recognition problem into a classification or re-

gression problem [17], it establishes a correspondence between the facial and head pose without relying on accurate facial feature point localization, enabling prediction of head pose with large deviations. Ruiz [33] proposed the Hopenet, which uses ResNet50 [34] as the backbone network to extract features and divides into three branches to jointly predict each angle. Each branch predicts the head pose angle by combining classification and regression objective functions. However, Hopenet performs coarse angle classification before aggregating regression, which introduces additional errors. To address this problem, Wang [35] proposed a hybrid coarse and fine classification framework introduced into this network, using more angle quantization units and other fine classifiers trained with auxiliary coarse units for better refinement of classification, which helps to reduce overfitting and improve the performance of prediction. Yang [21] proposed FSA-Net, which uses fine-grained structural mapping and scoring functions to filter important features and has a dual-stream multidimensional regression network based on the age classification algorithm SSR-Net [36]. By optimizing feature extraction and utilizing feature map aggregation of fine-grained structural mapping and scoring function to learn spatial relationships, it can estimate head pose without key-points information. Inspired by Hopenet, Zhou [37] proposed an end-to-end network model that changes the loss function and adapts to widely estimated training strategies to predict head pose angles across the entire range from a single image, which is the first method applicable to predicting head pose in the full range of head rotations. FDN [38] uses a feature decoupling model to explicitly identify the discriminative features for each angle by adaptively recalibrating the channel responses of each pose angle and suppressing less useful features. Zhu [39] proposed a hierarchical estimation method based on distinct network layers, gaining greater degrees of freedom in the angle estimation process. The LwPosr [40] uses a mixture of depth-separable convolution and Transformer [41] encoder layers with a dual-stream heterogeneous structure to extract features to provide fine-grained regression for predicting head pose. TriNet [42] considers orthogonal constraints on vectors to train the network, using three vectors to represent the head pose, and proposes Mean Absolute Error of Vectors to evaluate performance. There are also methods for estimating head pose based on multi-modal information. Gu [43] systematically analyzed the connection between Bayesian filtering and recurrent neural network (RNN) and use RNN to jointly estimate and track facial features in videos. Martin [44] proposed a method for head pose estimation on consumer depth cameras that combines head features and head model generation to build a detector that provides accurate results over a wide range of poses.

The deep learning-based models for head pose estimation employ an end-to-end recognition approach, which enables fast image processing, excellent high-dimensional feature extraction, strong generalization, and the extraction of the main features of head pose estimation tasks. However, this method faces the ambiguity of the Euler angle representation of head pose in a wide range of angles. Moreover, occlusion and multi-target detection significantly affect the accuracy of head pose estimation and result in the loss of target detection.

3. Method

In this section, we first formulate head pose estimation problem (Section 3.1). Then, we give an overview of the proposed network (Section 3.2). In Sections 3.3 and 3.4, we detail the feature extraction and feature mapping modules of the network. We illustrate the feature aggregation module and the use of SSR-Net to output the parameters predicted at each stage to obtain the final head pose estimation (Section 3.5). Finally, we optimize the regression loss function (Section 3.6).

3.1. Problem Formulation

Typically, head pose estimation can be represented as a regression problem-based image. We use a set of trained facial images $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ and the pose vector y_i for each image x_i, where i is the number of images. Each head pose vector y_i can

be subdivided into three angles: yaw, pitch, and roll. The goal of the regression task is to find a mapping function F and then use $\widetilde{y} = F(x)$ to predict the head pose angle of the input image. We find F by minimizing the Mean Absolute Error (MAE) between the predicted attitude and the ground truth attitude. Also, we use MAE to assess the performance of all methods, calculating MAE for yaw, pitch and roll separately and then averaging them for overall evaluation.

$$J(x) = \frac{1}{N}\sum_{n=1}^{N} \|\widetilde{y}_n - y_n\| \,, \tag{1}$$

where $\widetilde{y}_n = F(x_n)$ is the predicted pose for the trained image x_n. $J(x)$ is a function of the reduced MAE.

3.2. Overview of Proposed Network

The proposed network model is illustrated in Figure 1, which combines deep residual networks and improved capsule networks. It consists of three main components: the feature extraction module, feature mapping module, and feature aggregation module. For the task of feature extraction, both Hopenet [33] and TriNet [42] employ ResNet50 as the backbone network and have demonstrated good performance in various experiments. We employ the same backbone network to minimize unrelated variables and ensure a fair comparison.

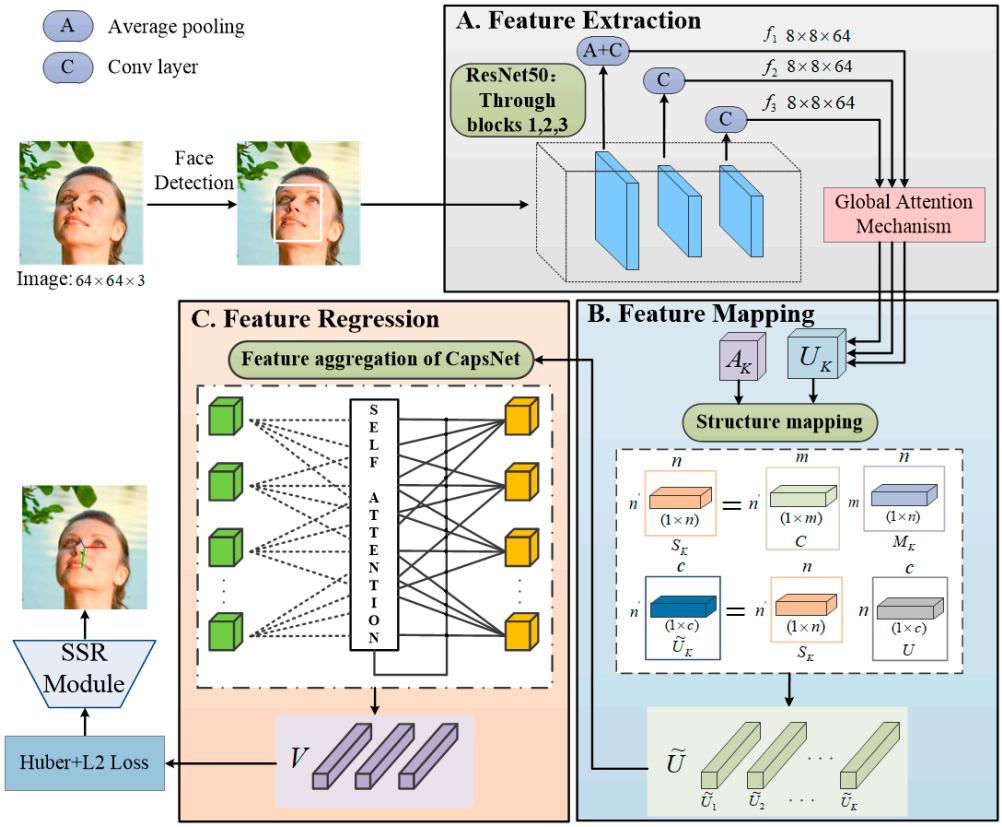

Figure 1. Overview of the proposed network.

Firstly, we employ multitask cascaded convolution neural network (MTCNN) [45] to detect the face region in the input image. Then, we feed the image into a feature extraction module, which adopts the Resnet50 with a multi-level output structure to extract feature maps from three stages and obtain spatial and semantic information at different levels. A global attention block (GAB) [46] is introduced in each stage to enhance the feature weight and extraction ability of key information and obtain more spatial information. The feature mapping module can obtain a more representative feature set. Next, we apply the improved capsule network to the feature aggregation module to obtain the final feature set.

We optimize the regression loss function to enhance the robustness of the model. Finally, we use SSR-Net to output the predicted parameters from each stage and obtain the final head pose estimation result.

3.3. Feature Extraction

3.3.1. Face Detection

Detecting faces and face alignment from images is challenging in varying unconstrained environments. Different lighting conditions, visual variations of faces and extreme head pose variations are the main challenges in correctly detecting faces from images [47]. Therefore, we use MTCNN as a face detector to detect faces and obtain their bounding box coordinates. It provides a real-time solution for detecting human heads across various scales and angles, even in the presence of complex and cluttered backgrounds. This capability is crucial for practical applications, where the quick and accurate detection of human heads is required. Figure 2 shows the architecture of the MTCNN. It is a network structure consisting of three layers of cascaded CNNs (P-Net, R-Net and O-Net). It utilizes bounding box regression and non-maximum suppression candidate filters to calibrate each layer of the network. The refined layer is more precise than the former, and the network parameters are trained to perform multiple tasks, resulting in a straightforward-to-advanced face detection process.

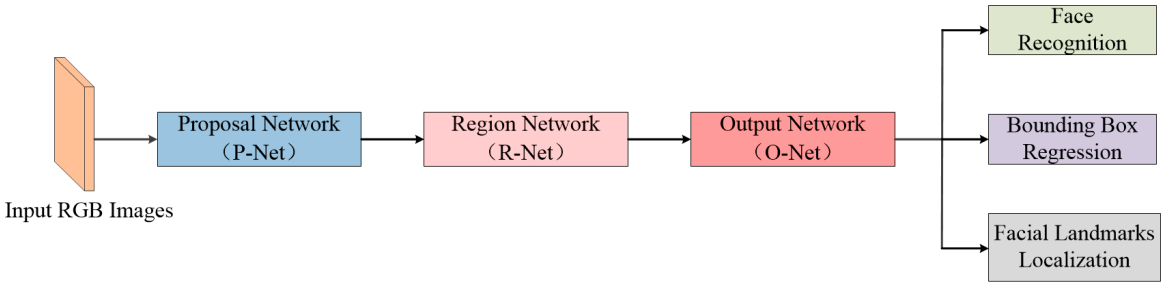

Figure 2. Architecture of MTCNN for face detection.

3.3.2. Global Attention Block

In deep convolution neural networks, attention mechanisms can refine feature maps to achieve better performance. To extract more effective features from the input face image, we design to introduce an efficient GAB that reduces information loss and amplifies cross-spatial-channel interactions across all three dimensions of importance. This module can improve the extraction efficiency of key features and the performance of deep neural networks. The structure diagram of GAM is shown in Figure 3. GAB captures global cross-space channel interactions, aiming to ensure the effectiveness and interactivity of feature information. It adopts the sequential channel-space attention mechanism from the convolution block attention module [48] and redesigns its sub-modules. The specific process can be represented by Equations (2) and (3). Given the input map $F_1 \in R^{C\times H\times W}$, the intermediate state F_2 and the output F_3 are defined as:

$$F_2 = M_c(F_1) \otimes F_1, \tag{2}$$

$$F_3 = M_s(F_2) \otimes F_2, \tag{3}$$

where M_c and M_s are the channel and spatial attention maps; $\otimes$ denotes element-based multiplication.

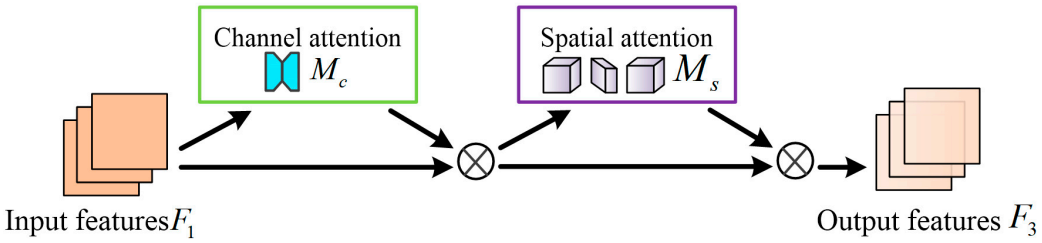

Figure 3. Architecture of the GAB.

The channel-attention sub-module uses 3D alignment to retain information, and then it uses a multi-layer perceptron to amplify the channel-space dependencies across dimensions. Finally, the original alignment is recovered, and a set of channel weights is generated after a sigmoid activation function to represent the weights of the feature mapping between channels. An enhanced feature map can be obtained by multiplying the weights with the input feature map. The channel attention sub-module is illustrated in Figure 4.

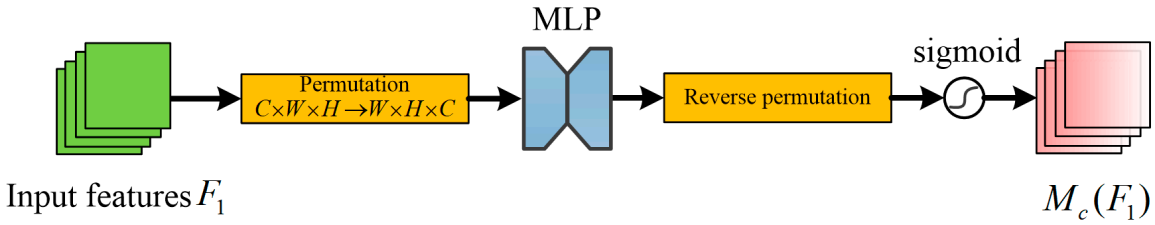

Figure 4. Architecture of the channel attention sub-module.

In the spatial attention sub-module, two convolution layers are used for spatial information fusion and the same reduction ratio as in the channel attention sub-module to focus spatial information. Also, pooling is removed to further preserve feature mapping. The spatial attention sub-module is illustrated in Figure 5.

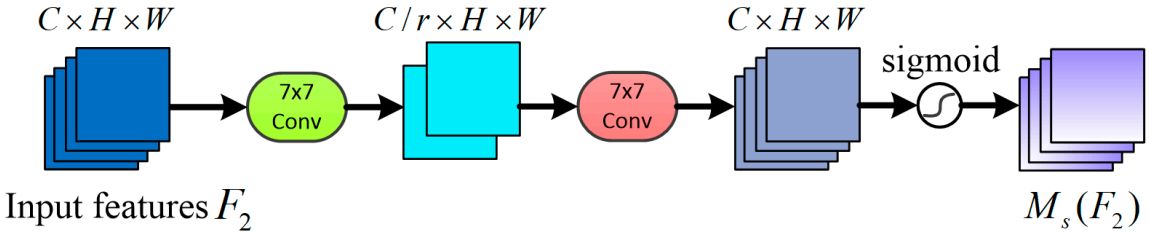

Figure 5. Architecture of the spatial attention sub-module.

3.4. Feature Mapping

After the above feature extraction module, we obtain the feature maps U_k, whose dimension is $w \times h \times c$. Figure 6 illustrates the structure of the feature mapping module. To obtain smaller and representative feature maps, a scoring function $\phi(u)$ is used to measure their importance. Similar to [21], we force the fine-grained mapping module. The function $\phi(u)$ includes 1×1 convolution, variance, and uniform to facilitate spatial grouping. For each feature map U_k, we obtain an attention map A_k by Equation (4).

$$A_k = \phi(U_k(i,j)). \tag{4}$$

The next step is to perform fine-grained structure mapping to obtain more representative features $\tilde{U}$. Figure 6 illustrates the process. All feature maps are first flattened into a 2D matrix U, where $U \in R^{n\times c}$ and $n = w \times h \times k$. The matrix U contains all the pixel-level

features in all phase feature maps. For the k-th stage, we design a mapping S_k to extract n' representative features $\widetilde{U}_k$ by Equation (5).

$$\widetilde{U}_k = S_k \times U, \tag{5}$$

where $S_k \in R^{n' \times n}$ and $\widetilde{U}_k \in R^{n' \times c}$. In other words, we obtain a representative feature by a linear combination of pixel-level features. Mapping S_k is a linear transformation that performs linear dimensionality reduction by a weighted average of all pixel-level features. The mapping S_k is the multiplication of C and M_k. The maps M_k and C are formed as follows:

$$M_k = \sigma(f_M(A_k)), \tag{6}$$

$$C = \sigma(f_C(A)), \tag{7}$$

where $C \in R^{n' \times m}$ and $M_k \in R^{m \times n}$ is the sigmoid function; f_M and f_C are two different functions defined by fully connected layers. $A = [A_1, A_2, \ldots, A_k]$ is the concatenation of all attentive maps. Finally, all features $\widetilde{U}_k$ are connected to form a final set of representative features $\widetilde{U} = \left[\widetilde{U}_1, \widetilde{U}_2, \ldots, \widetilde{U}_k\right]$, where $\widetilde{U} \in R^{(n' \times k) \times c}$.

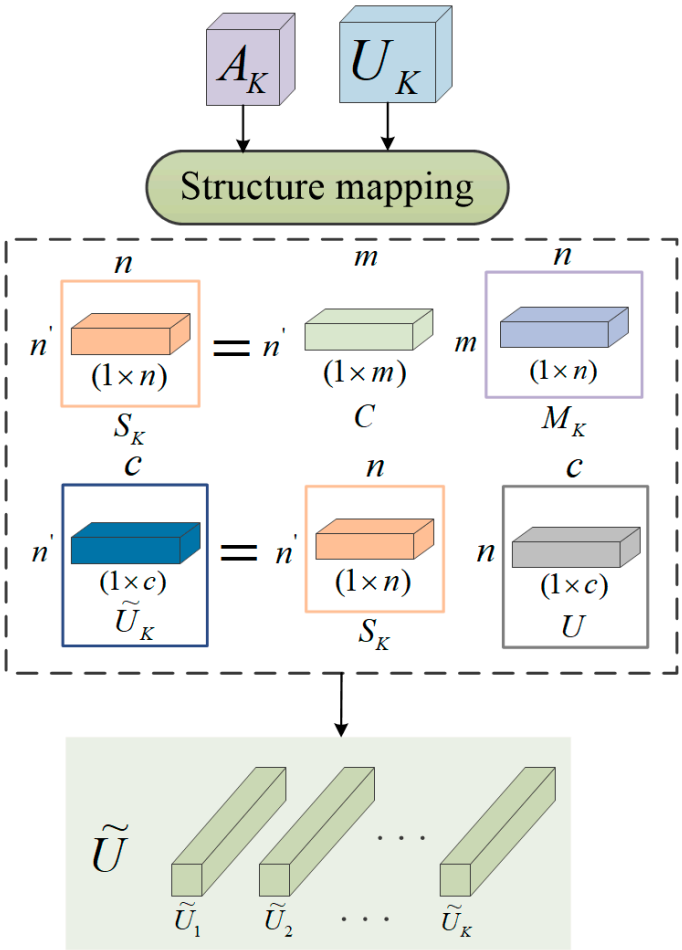

Figure 6. Architecture of the feature mapping model.

3.5. Feature Aggregation

Regarding the aggregation module, we first consider CNNs because their convolutional structures can effectively capture existing features. However, CNNs rely on large amounts of data and layers with feature mappings to complete learning and updates, which may not be very efficient. When capturing relationships between feature attributes,

CNNs may cause feature detectors to lose the precise target information from input images. Capsule networks provide a good solution, as they can extend current convolutional networks to efficiently encode all affine transformations of features and have better generalization capabilities.

To overcome the limitations of the CNN method, we adopt the capsule network for feature aggregation. However, the capsule network requires training a large number of parameters, leading to low utilization efficiency and insufficient inherent generalization ability in expressing feature transformations. To address this problem, we use the linear combination and self-attention routing mechanism (SARM) [18]. Utilizing the method that can acquire more comprehensive features can enhance the network's capability in facial feature extraction, thereby decreasing the effect of absent facial feature information on the forecasting outcomes. At the same time, it also effectively reduces the number of capsules, enhances the generalization ability of the capsule network and improves the network's recognition performance for head posture.

As shown in Figure 7, the mechanism is similar to a fully connected network with additional branches introduced by the self-attention algorithm. In fact, the upper-level capsules receive the total input as a weighted sum of the "predicted vectors" from the lower-level capsules. The weight matrix is obtained by matrix multiplication for each capsule. The tensor that contains all the weight matrices is then embedded into the affine transformations between adjacent capsules, allowing the lower-level capsules to predict the properties of all upper-level capsules. This tensor mainly consists of the log prior matrix and the coupling coefficient matrix. The log prior matrix includes all the weights that are learned discriminatively with other weights, which helps to establish more closely related capsules. The coupling coefficient matrix is a matrix that contains all the coupling coefficients generated by the self-attention algorithm. The self-attention routing dynamically assigns the detected shape to the represented entity. Therefore, the lower-level capsules can efficiently aggregate features by predicting the upper-level capsules through this mechanism, reducing the number of capsules and trainable parameters, and obtaining better results.

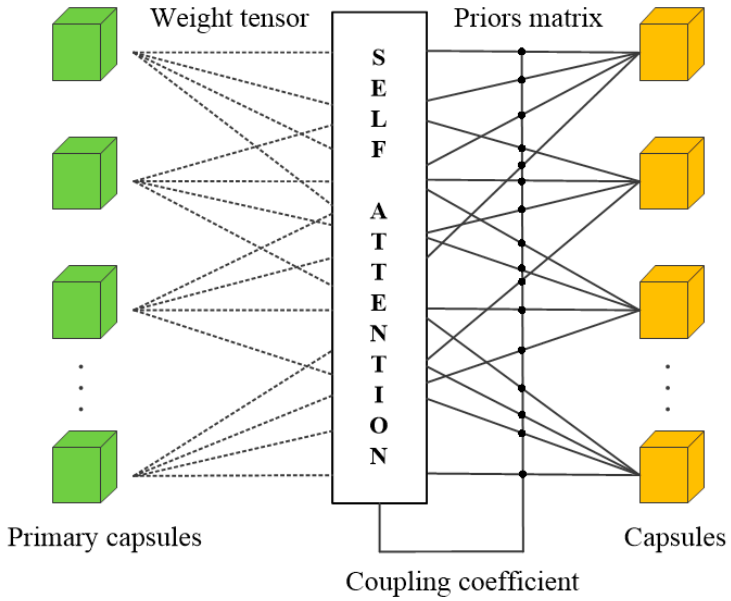

Figure 7. Architecture of the SARM.

3.6. Optimization of Loss Functions

For the loss function of the regression model, TriNet [42] employs the Mean Square Error (MSE) loss function for both regression and orthogonal losses. The gradient of MSE decreases as the error decreases, which is beneficial to the convergence of the model. However, the MSE loss function is susceptible to the influence of outliers, causing the

regression line to shift towards the outlier data points. FSA-Net [21] uses the MAE loss function, which has a stable gradient and does not cause gradient explosion. Moreover, it is less affected by outliers and has greater inclusivity, which allows the fitted line to better characterize the distribution of normal data. In most cases, the gradient of MAE remains constant, which implies that even for small losses, the gradient still holds a significant value. This characteristic could impede function convergence and the learning of the model.

To address this problem, we propose a novel regression loss function L_{hul} that incorporates Huber loss [49] and L2 regularization loss. To the best of our knowledge, this is the first application of Huber loss to head pose estimation. Huber loss is a loss function that combines MSE and MAE, with a hyperparameter α that determines the weight between the two components. When $|g - f(t)| \leq \alpha$, Huber loss becomes MSE. When $|g - f(t)| > \alpha$, Huber loss approximates to MAE. Huber loss, therefore, offers the advantages of both MSE and MAE, reducing the problem of sensitivity to outliers. Also, we use L2 regularization loss to ensure the generalization of the model, as described in Equation (8).

$$L_{hul} = \begin{cases} \frac{1}{2}(g - f(t))^2 + \lambda \|w\|_2{}^2, |g - f(t)| \leq \alpha \\ \alpha |g - f(t)| - \frac{1}{2}\alpha^2 + \lambda \|w\|_2{}^2, |g - f(t)| > \alpha \end{cases}, \tag{8}$$

where g is the sample true pose, $f(t)$ is the prediction pose, λ is the regularization factor, and w is the regularization parameter.

4. Experiments

In this section, we provide a detailed account of the experimental procedure and results. The first part describes the experimental setup, while the second part introduces the dataset and evaluation metrics used in the experiments. The third part presents experimental results to test the effectiveness of the proposed network. Finally, we compare the experimental results with those of typical head pose estimation algorithms to further validate the performance of our method.

4.1. Experimental Implementation

The experiments are performed on a computer with an Intel Xeon Gold 5118 CPU and Nvidia RTX 5000 GPU. We use Keras with TensorFlow 1.10.0 backend to implement the proposed network. To make a fair comparison, we apply random cropping and random scaling to training images by following the data augmentation strategies in training from FSA-Net [21] and TriNet [42]. We apply Adam [50] as the optimizer for training with the initial learning rate 0.001. We use 100 epochs to train the network and the learning rate is reduced by a factor of 0.1 every 30 epochs. To improve the processing of blurred and enlarged images, random cropping and random scaling are used on the training images to enhance the training data.

4.2. Datasets and Experimental Protocols

In Figure 8, the experiments used three datasets for head pose estimation: 300W-LP [26], AFLW2000 [51], and BIWI [52].

300W-LP: The 300W-LP dataset is a combined dataset containing eight small datasets. It has a total of 122,450 images and contains samples of large-angle head poses, and the images are labeled with the Euler angles rotated during the image processing. This dataset is currently the largest publicly available dataset in the field of head pose estimation and has been chosen for the training of head pose estimation models proposed in recent years.

AFLW2000: The AFLW2000 dataset consists of 2000 images selected from the AFLW dataset, with most of its sample being portraits of people in natural scenes. The images in the AFLW2000 dataset include variations in the pose of people in different scenes and brightness, covering people of different ages and ethnicities. It has pretty accurate face pose annotation providing ground-truth 3D faces and the corresponding 68 landmarks.

Therefore, we test the model using the AFLW2000 dataset to verify the generalization capability of the model.

BIWI: The BIWI dataset is a publicly available dataset commonly used in the field of head pose estimation, with approximately 15,000 images. The dataset contains 24 videos of head poses from 20 test subjects (6 females and 14 males). In addition to RGB frames, the dataset also provides the depth image for each frame. Ground-truth is provided in the form of the 3D location of the head and its rotations.

Figure 8. Sample images from the three datasets.

For comparison with state-of-the-art methods, we follow the same training and testing setup as mentioned in Hopenet [33], FSA-Net [21], and TriNet [42]. For training and testing on these datasets, we implement our experiments in two common protocols:

(1) In our first protocol, we follow the convention by using the 300W-LP dataset for training and AFLW2000 and BIWI datasets for testing. When evaluating on the BIWI dataset, we do not use tracking and only consider using MTCNN [45] face detection samples whose rotation angles are in the range of $[-99^{\circ}, +99^{\circ}]$ to keep consistent with the strategies used by Hopenet [33], FSA-Net [21] and TriNet [42]. Also, we compare several state-of-the-art landmark-based pose estimation methods using this protocol.

(2) For the second protocol, we follow the convention by FSA-Net [21] and randomly split the BIWI dataset in a ratio of 7:3 for training and testing, respectively. The train set is not crossed with the test set. MTCNN [45] uses experience tracking technology to detect faces in the BIWI datasets, avoiding the failure of face detection. This protocol is used by several pose estimation methods such as RGB, depth, and time, whereas our method uses only a single RGB frame.

In all the experiments above, we evaluate the performance of all methods using MAE. For each method, MAE for yaw, pitch, and roll is separately calculated and then taking the average for overall evaluation.

4.3. Experiment Results

We explore the performance variation of the model using different loss functions, namely MAE, Huber loss and L_{hul} loss. Table 1 shows the comparative results of the three different loss functions for protocol 1.

We test each of the three loss functions by incorporating them into the improved model, where "$\times$" indicates that the improved model is not used and "$\surd$" indicates that the improved model is used. The result shows that the improved model can improve performance by incorporating different loss functions. We observe that there is little difference in performance between both using MAE and Huber loss. We analyze that the use of Huber loss leads to overfitting of the model. We propose that L_{hul} can be a good solution to the problem and greatly improve the performance of the model.

Table 1. Performance comparison between different loss functions on the AFLW2000 and BIWI datasets. All are trained on the 300W-LP dataset. The bolded data is the most effective.

Loss	Improved	AFLW2000				BIWI			
		Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
MAE	×	4.49	6.22	4.64	5.12	4.39	5.36	2.76	4.17
MAE	√	4.00	6.17	4.27	4.81	4.36	4.89	2.64	4.05
Huber loss	×	4.12	6.37	4.58	5.02	4.51	5.16	2.71	4.13
Huber loss	√	3.94	6.13	4.27	4.78	4.42	4.90	2.59	3.97
L_{hul} loss	×	4.03	6.21	4.38	4.87	4.34	5.02	2.70	4.02
L_{hul} loss	√	**3.91**	**5.78**	**4.11**	**4.60**	**4.25**	**4.96**	**2.57**	**3.93**

4.3.1. Results with Protocol 1

In protocol 1, the features of the training and test datasets are completely different, with the training dataset being synthetic and the test dataset being real. The landmark-free approach can better accommodate the domain differences between training and testing. As a result, the landmark-free approach performs better than the landmark-based approach on both the AFLW2000 and BIWI datasets.

Tables 2 and 3 compare the proposed method with the state-of-the-art methods on the AFLM2000 and BIWI datasets, respectively. As shown in Table 2, the proposed method realizes the best performance and attains the minimum error on yaw when tested on the AFLW2000 dataset. Pitch and roll angle errors are somewhat higher than TriNet, but an average deviation angle error reduction of 0.07 compared to TriNet. Furthermore, in Table 3, the experimental results on the BIWI dataset are shown. The proposed method reaches minimum error on a roll, and other indicators are also in the upper middle position with an average deviation angle error of 3.93.

Table 2. Comparisons with the state-of-the-art methods on the AFLW2000 dataset. All are trained on the 300W-LP dataset. The bolded data is the most effective.

Method	Yaw	Pitch	Roll	MAE
Dlib [25]	23.10	13.60	10.50	15.80
FAN [30]	6.36	12.3	8.71	9.12
3DDFA [26]	5.40	8.53	8.25	7.39
Hopenet [33]	6.47	6.56	5.44	6.16
SSR-Net-MD [36]	5.14	7.09	5.89	6.01
FSA-Net [21]	4.50	6.08	4.64	5.07
WHENet [37]	5.11	6.24	4.92	5.42
TriNet [42]	4.20	**5.77**	**4.04**	4.67
LwPosr [40]	4.80	6.38	4.88	5.35
Ours	**3.91**	5.78	4.11	**4.60**

Table 3. Comparisons with the state-of-the-art methods on the BIWI dataset. All are trained on the 300W-LP dataset. The bolded data is the most effective.

Method	Yaw	Pitch	Roll	MAE
Dlib [25]	16.80	13.80	6.19	12.2
FAN [30]	8.53	7.48	7.63	7.89
3DDFA [26]	36.20	12.30	8.78	19.10
Hopenet [33]	4.81	6.61	3.27	4.90
KEPLER [31]	8.80	17.3	16.2	13.9
SSR-Net-MD [36]	4.49	6.31	3.61	4.65
FSA-Net [21]	4.27	4.96	2.76	4.00
TriNet [42]	**3.05**	**4.76**	4.11	3.97
LwPosr [40]	4.11	4.87	3.19	4.05
Ours	4.25	4.96	**2.57**	**3.93**

4.3.2. Results with Protocol 2

Table 4 compares the performance of other state-of-the-art methods on the BIWI dataset, where 70% and 30% of the data are randomly splatted for training and testing, without crossover. However, TriNet [42] applies a 3-fold cross validation on BIWI dataset in this protocol. It split the dataset into three groups and ensures that the images of one person should appear in the same group. For a fair comparison, we return the open-sourced models and measure the results of MAE under the same experimental protocol. The BIWI dataset contains multiple modes of information, and in addition to RGB information, depth or temporal information can be used to improve performance. The finding of methods based on multi-modal information is derived from FSA-Net [21]. The results show that our model does not perform as well as methods using multiple modalities on pitch and roll, but it only uses a single RGB frame and outperforms all other methods in its peer group. In addition, our method achieves the best yaw angle estimation and overall performance, even outperforming methods that utilize multiple modalities.

Table 4. Comparisons with the state-of-the-art methods on the BIWI dataset. 70% of videos are trained and 30% for testing. The bolded data is the most effective.

Method	Input	Yaw	Pitch	Roll	MAE
SSR-Net-MD [36]	RGB	4.24	4.35	4.19	4.26
FSA-Net [21]	RGB	2.89	4.29	3.60	3.60
TriNet [42]	RGB	3.18	3.57	2.85	3.20
VGG16+RNN [43]	RGB + Time	3.14	3.48	**2.60**	3.07
Martin [44]	RGB + Depth	3.60	**2.50**	**2.60**	2.90
Ours	RGB	**2.37**	3.14	2.83	**2.78**

4.3.3. Results with Model Size and Computation Time

To fully understand the performance of the proposed network, we compare the proposed method with other state-of-the-art methods in terms of model size, parameters, and computation time. For a fair comparison, all networks are tested on the same target platform. Table 5 shows that our method achieves a computation time of about 12 fps, which outperforms other state-of-the-art methods. Our method is larger than FSA-Net in terms of model size and parameters but performs better in head pose estimation. Moreover, compared to Hopenet [33] and TriNet [42], our method applies the same backbone network but has a smaller model size and parameters.

Table 5. Comparisons with the state-of-the-art methods in terms of model size, parameters, and FPS. The bolded data is the most effective.

Method	Image Size	Model Size (MB)	Parameters (M)	FPS
Hopenet [33]	224 × 224	95.9	23.92	9
FSA-Net [21]	64 × 64	**5.1**	**1.17**	6
TriNet [42]	64 × 64	27.96	1.95	10
Ours	64 × 64	21.3	1.68	**12**

4.4. Visualization

Figure 9 presents the visualizations of feature maps extracted by the feature extraction module at different stages. First, coarse features are extracted from the input image. Then, through adjusting channel sizes and performing consecutive convolution and pooling operations, deeper-level features are extracted from the outputs of the previous stage. Finally, redundant information is removed from the features, while preserving the relevant information to enhance the efficiency of the network. The results demonstrate that the proposed method effectively captures representative features for occluded and angle-deflected images.

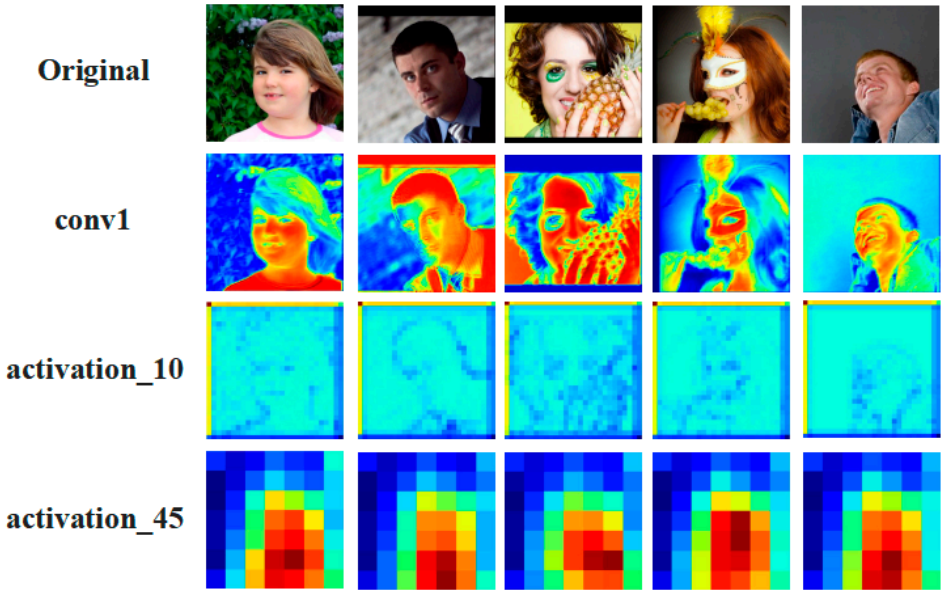

Figure 9. The feature map visualizations of head pose with occlusion and significant angle deflection on the AFLW2000 dataset.

To further validate the effectiveness of our proposed method, we present a visualization comparison between different methods. Figure 10 shows the head pose estimation results of the proposed method and FSA-Net on images with significant angle deviations. Three different colored lines are used to represent pitch, yaw, and roll directions, respectively, making the head pose estimation results visualizable. Specifically, the blue line indicates the direction of the face, the green line indicates the direction of the bottom, and the red line indicates the direction of the side.

Figure 10. Estimation of head pose with significant angle deflection on the AFLW2000 dataset.

In this study, we propose a method for head pose estimation that shares a similar feature mapping module with FSA-Net. To demonstrate the effectiveness of our proposed method, we select a subset of images with significant occlusions from the AFLW2000 dataset. Figure 11 shows some challenging examples. The visualization analysis results indicate that our proposed method is closer to the ground truth labels in cases of significant angle deviation and occlusions compared to FSA-Net.

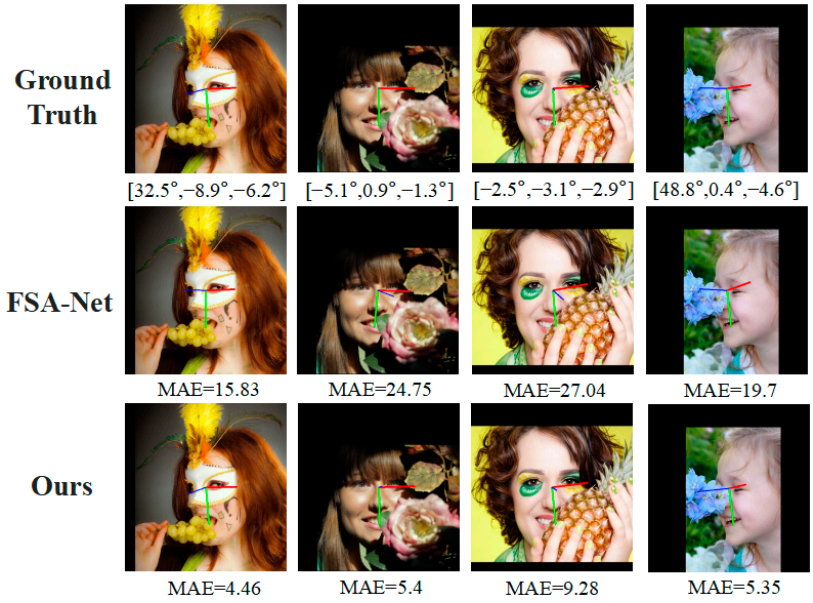

Figure 11. Estimation of head pose with occlusion on the AFLW2000 dataset.

In addition, to strengthen the persuasive power of our experiments, we refer to the visualization results of Zhu [39]. Based on their work, we compare our proposed method with other methods. The results in Figure 12 show that our method achieves the best performance.

Figure 12. Various methods for estimating head pose sample images on the AFLW2000 dataset. The input images and results of other methods from (Zhu 2022 [39]).

4.5. Ablation Study

In this section, we conducted ablation experiments to investigate the impact of different modules (feature extraction module, feature aggregation module, and regression loss function) on the performance of our proposed model. The proposed method was trained on the 300W-LP dataset and tested on the AFLW2000 and BIWI datasets for the three

modules. We then used 70% of the BIWI dataset as a training set and 30% as a test set. The experimental results for the different datasets are shown in Tables 6 and 7, respectively.

Table 6. Ablation study for different feature extraction modules (with/without GAB) and capsule network (with/without SARM) and loss function (with/without L_{hul} loss). All are trained on the 300W-LP dataset. The bolded data is the most effective.

GAB	SARM	L_{hul} Loss	AFLW2000				BIWI			
			Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
×	×	×	4.55	6.24	4.72	5.17	4.37	5.46	2.89	4.24
√	×	×	4.31	6.32	4.59	5.07	4.33	5.41	2.94	4.20
×	√	×	4.43	6.15	4.45	5.01	4.35	5.24	2.80	4.13
×	×	√	4.47	6.28	4.52	5.09	4.40	5.39	2.69	4.16
√	√	×	4.05	5.87	4.33	4.75	4.32	5.08	2.60	4.00
√	×	√	4.14	6.11	4.42	4.89	4.30	5.25	2.72	4.09
×	√	√	4.20	6.06	4.23	4.83	4.36	5.21	2.61	4.06
√	√	√	3.91	5.78	4.11	**4.60**	4.25	4.96	2.57	**3.93**

Table 7. Ablation study for different feature extraction modules (with/without GAB) and capsule network (with/without SARM) and loss function (with/without L_{hul} loss). All are trained on the BIWI dataset (70%). The bolded data is the most effective.

GAB	SARM	L_{hul} Loss	BIWI			
			Yaw	Pitch	Roll	MAE
×	×	×	2.96	4.43	3.80	3.73
√	×	×	2.91	3.45	3.21	3.19
×	√	×	2.56	3.48	3.19	3.07
×	×	√	2.73	3.53	3.07	3.11
√	√	×	2.45	3.25	2.90	2.87
√	×	√	2.60	3.41	2.91	2.97
×	√	√	2.39	3.27	2.87	2.84
√	√	√	2.37	3.14	2.83	**2.78**

Table 6 shows the results of the ablation experiments for the proposed method on the AFLW2000 and BIWI datasets. The baseline model achieves an MAE of 5.17 on the AFLW2000 dataset and an MAE of 4.24 on the BIWI dataset when the proposed modules are not used. From the results of adding each of the three modules separately, it can be seen that each module improves the performance of the model, with SARM having the best performance on the AFLW2000 dataset. We analyzed that this mechanism improves the performance and generalization ability of the capsule network by reducing the number of capsules and trainable parameters, effectively avoiding the problem of spatial structural information loss. When all modules are combined, the proposed model achieves an MAE of 4.60 on the AFLW2000 dataset and an MAE of 3.93 on the BIWI dataset. This indicates that all three proposed modules are effective.

Table 7 presents the ablation study results of our proposed model based on RGB information on the BIWI dataset, which includes depth information of face images and head poses with different angles. Compared to the baseline model, our proposed method reduces the MAE by 0.95, indicating its suitability for head poses with different angle ranges. Figure 13 provides a detailed comparison of yaw, pitch, and roll angles under several settings of the ablation study. The best experimental results are obtained when all three modules are combined, demonstrating that these modules together enhance the accuracy of head pose estimation, significantly improving the performance of head pose estimation algorithms, and thus validating the effectiveness of our proposed method.

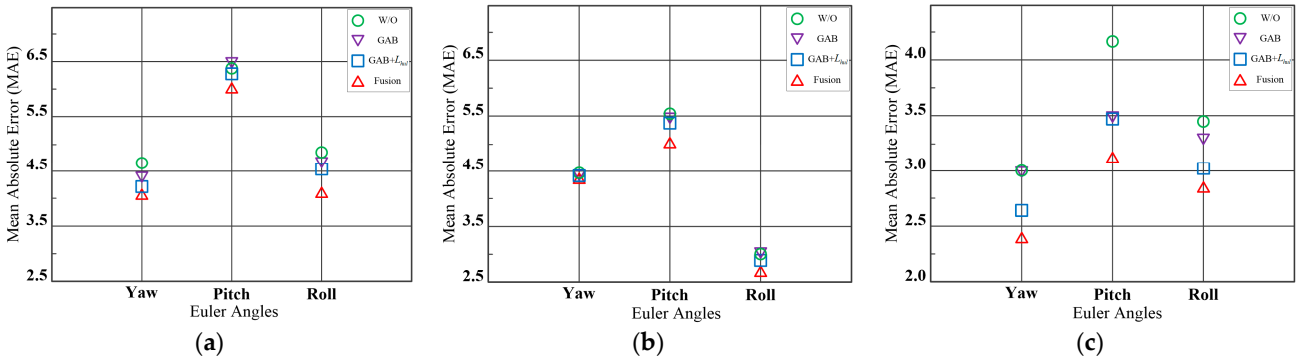

Figure 13. Comparison of the MAE of various components at various angles under protocol 1 and protocol 2. We divide the components of the proposed method into three parts, GAB, GAB fusion L_{hul} loss, and all modules fused. "w/o" denotes without the proposed method. (**a**) AFLW2000 (protocol 1); (**b**) BIWI (protocol 1); (**c**) BIWI (protocol 2).

5. Conclusions

To avoid loss of spatial information and improve the estimation accuracy of head pose under occlusion, in this paper, we propose a head pose estimation model based on a fusion of a convolutional neural network and capsule network. We adopt a feature extraction network with a multi-level output structure and introduce GAB to enhance the spatial weight of feature extraction, which can obtain more spatial information. The capsule network can retain and transmit spatial information. We improve the capsule network with SARM to significantly reduce the number of capsules and trainable parameters, effectively improving the model's performance and solving the problem of loss of spatial structural information. In addition, to enhance the robustness of the model, we have utilized Huber loss for head pose estimation for the first time, which has better performance compared to methods based on multiple loss functions. The collaborative fusion of convolutional neural networks and capsule networks facilitates the extraction and preservation of both spatial and semantic features, resulting in better detection of head pose.

This study conducts experiments on AFLW2000 and BIWI datasets, comparing the proposed method with previous methods. The results demonstrate that the proposed model exhibits more advanced performance in cases of significant angle deviations and occlusions. In future work, we will primarily focus on real-time head pose detection for analysis in social interactions. To achieve this, we will investigate lightweight network architectures to further improve computational speed and optimize the model parameters. Additionally, we intend to incorporate multimodal information, such as depth and temporal cues, for head pose estimation, with the aim of enhancing the performance of the network.

Author Contributions: Conceptualization, R.Z.; methodology, R.Z., K.L. and Z.L.; software, L.Y.; data collection and pre-processing, R.Z., K.L. and Z.L.; validation, R.Z. and K.L.; investigation, Z.L.; writing—original draft preparation, R.Z.; writing—review and editing, L.H. and H.W.; visualization, R.Z.; supervision, L.H. and H.W.; project administration, L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research is sponsored by the National Natural Science Foundation of China (62063033), the Key R&D Program of Xinjiang Uygur Autonomous Region (2022B01050-2) and the Natural Science Foundation of Xinjiang Uygur Autonomous Region (2022D01C392).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: We use three publicly available datasets, 300W-LP, AFLW2000, and BIWI in our experiments. Their links are as follows: 300WLP: http://www.cbsr.ia.ac.cn/users/xiang yuzhu/projects/3DDFA/main.htm; AFLW2000: http://www.cbsr.ia.ac.cn/users/xiangyuzhu/prjects /3DDFA/main.htm; BIWI: https://data.vision.ee.ethz.ch/cvl/gfanelli/head_pose/head_forest.html (all of the above datasets accessed on 23 April 2023).

Acknowledgments: We are very grateful to anonymous reviewers for their valuable and insightful suggestions on the original manuscript.

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

References

1. Moller, R.; Furnari, A.; Battiato, S. A survey on human-aware robot navigation. *Robot. Auton. Syst.* **2021**, *145*, 103837. [CrossRef]
2. Murphy-Chutorian, E.; Trivedi, M.M. Head pose estimation in computer vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 607–626. [CrossRef] [PubMed]
3. Jie, S.; Lu, S. An improved single shot multibox for video-rate head pose prediction. *IEEE Sens. J.* **2020**, *20*, 12326–12333.
4. Yining, L.; Liang, W.; Fang, X.; Yibiao, Z.; Lap-Fai, Y. Synthesizing Personalized Training Programs for Improving Driving Habits via Virtual Reality. In Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Tuebingen/Reutlingen, Germany, 18–22 March 2018; pp. 297–304.
5. Ye, M.; Zhang, W.; Cao, P. Driver fatigue detection based on residual channel attention network and head pose estimation. *Appl. Sci.* **2021**, *11*, 9195. [CrossRef]
6. Fan, Z.; Li, X.; Li, Y. Multi-Agent Deep Reinforcement Learning for Online 3D Human Poses Estimation. *Remote Sens.* **2021**, *13*, 3995. [CrossRef]
7. Murphy-Chutorian, E.; Trivedi, M.M. Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness. *IEEE Trans. Intell. Transp. Syst.* **2010**, *11*, 300–311. [CrossRef]
8. Vankayalapati, H.D.; Kuchibhotla, S.; Chadalavada, M.S.K. A Novel Zernike Moment-Based Real-Time Head Pose and Gaze Estimation Framework for Accuracy-Sensitive Applications. *Sensors* **2022**, *22*, 8449. [CrossRef]
9. Qi, S.; Wang, W.; Jia, B. Learning human-object interactions by graph parsing neural networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 401–417.
10. Wang, K.; Zhao, R.; Ji, Q. Human computer interaction with head pose, eye gaze and body gestures. In Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG), Xi'an, China, 15–19 May 2018; p. 789.
11. Sankaranarayanan, K.; Chang, M.C.; Krahnstoever, N. Tracking gaze direction from far-field surveillance cameras. In Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV), Kona, HI, USA, 5–7 January 2011; pp. 519–526.
12. Chen, C.W.; Aghajan, H. Multiview social behavior analysis in work environments. In Proceedings of the 5th ACM/IEEE International Conference on Distributed Smart Cameras, Ghent, Belgium, 22–25 August 2011; pp. 1–6.
13. Yunjuan, H.; Li, F. Isospectral Manifold Learning Algorithm. *J. Softw.* **2013**, *24*, 2656–2666.
14. Wu, J.; Shang, Z.; Wang, K. Partially Occluded Head Posture Estimation for 2D Images using Pyramid HoG Features. In Proceedings of the 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shanghai, China, 8–12 July 2019; pp. 507–512.
15. Yujia, W.; Wei, L.; Jianbing, S.; Yunde, J. A deep Coarse-to-Fine network for head pose estimation from synthetic data. *Pattern Recognit.* **2019**, *94*, 196–206.
16. Junliang, X.; Zhiheng, N.; Junshi, H. Towards robust and accurate multi-view and partially-occluded face alignment. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 987–1001.
17. Bisogni, C.; Nappi, M.; Pero, C.; Ricciardi, S. FASHE: A FrActal Based Strategy for Head Pose Estimation. *IEEE Trans. Image Process.* **2021**, *30*, 3192–3203. [CrossRef]
18. Mazzia, V.; Salvetti, F.; Chiaberge, M. Efficient-capsnet: Capsule network with self-attention routing. *Sci. Rep.* **2021**, *11*, 14634. [CrossRef]
19. Hinton, G.E.; Krizhevsky, A.; Wang, S.D. Transforming auto-encoders. In *Artificial Neural Networks and Machine Learning–ICANN, Proceedings of the 21st International Conference on Artificial Neural Networks, Espoo, Finland, 14–17 June 2011*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 44–51.
20. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic routing between capsules. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3856–3866.
21. Yang, T.; Chen, Y.; Lin, Y.; Chuang, Y. FSA-Net: Learning fine-grained structure aggregation for head pose estimation from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1087–1096.
22. Chang, F.J.; Tran, A.T.; Hassner, T. Expnet: Landmark-free, deep, 3d facial expressions. In Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG), Xi'an, China, 15–19 May 2018; pp. 122–129.
23. Liu, L.; Ke, Z.; Huo, J. Head pose estimation through keypoints matching between reconstructed 3D face model and 2D image. *Sensors* **2021**, *21*, 1841. [CrossRef] [PubMed]
24. Li, D.; Pedrycz, W. A central profile-based 3D face pose estimation. *Pattern Recognit.* **2014**, *47*, 525–534. [CrossRef]
25. Kazemi, V.; Sullivan, J. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 1867–1874.
26. Zhu, X.; Lei, Z.; Liu, X.; Shi, H.; Li, S.Z. Face alignment across large poses: A 3D solution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2–30 June 2016; pp. 146–155.
27. Nikolaidis, A.; Pitas, I. Facial feature extraction and pose determination. *Pattern Recognit.* **2000**, *33*, 1783–1791. [CrossRef]
28. Illingworth, J.; Kittler, J. The adaptive Hough transform. *IEEE Trans. Pattern Anal. Mach. Intell.* **1987**, *9*, 690–698. [CrossRef]

29. Narayanan, A.; Kaimal, R.M.; Bijlani, K. Estimation of driver head yaw angle using a geometric model. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 3446–3460. [CrossRef]
30. Bulat, A.; Tzimiropoulos, G. How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks). In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1021–1030.
31. Kumar, A.; Alavi, A.; Chellappa, R. KEPLER: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. In Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG), Washington, DC, USA, 30 May–3 June 2017; pp. 258–265.
32. Wang, Q.; Lei, H.; Qian, W. Siamese PointNet: 3D Head Pose Estimation with Local Feature Descriptor. *Electronics* **2023**, *12*, 1194. [CrossRef]
33. Ruiz, N.; Chong, E.; Rehg, J.M. Fine-grained head pose estimation without keypoints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 215501–215509.
34. He, K.; Zhang, X.; Ren, S. 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.
35. Wang, H.; Chen, Z.; Zhou, Y. Hybrid coarse-fine classification for head pose estimation. *arXiv* **2019**, arXiv:1901.06778.
36. Yang, T.; Huang, H.; Lin, Y.; Hsiu, P.; Chuang, Y. SSR-Net: A compact soft stagewise regression network for age estimation. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, 13–19 July 2018; pp. 1078–1084.
37. Zhou, Y.; Gregson, J. WHEnet: Real-time fine-grained estimation for wide range head pose. *arXiv* **2020**, arXiv:2005.10353.
38. Zhang, H.; Wang, M.; Liu, Y.; Yuan, Y. FDN: Feature decoupling network for head pose estimation. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), New York, NY, USA, 7–12 February 2020; pp. 12789–12796.
39. Zhu, X.; Yang, Q.; Zhao, L. An Improved Tiered Head Pose Estimation Network with Self-Adjust Loss Function. *Entropy* **2022**, *24*, 974. [CrossRef] [PubMed]
40. Dhingra, N. Lwposr: Lightweight efficient fine grained head pose estimation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022; pp. 1495–1505.
41. Dhingra, N. HeadPosr: End-to-end Trainable Head Pose Estimation using Transformer Encoders. In Proceedings of the 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG), Jodhpur, India, 15–18 December 2021; pp. 1–8.
42. Cao, Z.; Chu, Z.; Liu, D.; Chen, Y. A vector-based representation to enhance head pose estimation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 1187–1196.
43. Jiawei, G.; Xiaodong, Y. Dynamic Facial Analysis: From Bayesian Filtering to Recurrent Neural Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1531–1540.
44. Martin, M.; Van De Camp, F.; Stiefelhagen, R. Real time head model creation and head pose estimation on consumer depth cameras. In Proceedings of the 2nd International Conference on 3D Vision (3DV), Tokyo, Japan, 8–11 December 2014; pp. 641–648.
45. Zhang, K.; Zhang, Z.; Li, Z. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
46. Liu, Y.; Shao, Z.; Hoffmann, N. Global attention mechanism: Retain information to enhance channel-spatial interactions. *arXiv* **2021**, arXiv:2112.05561.
47. Joshi, M.; Pant, D.R.; Karn, R.R. Meta-Learning, Fast Adaptation, and Latent Representation for Head Pose Estimation. In Proceedings of the 31st Conference of Open Innovations Association (FRUCT), Helsinki, Finland, 27–29 April 2022; pp. 71–78.
48. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In *Lecture Notes in Computer Science, Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–19.
49. Huber, P.J. Robust estimation of a location parameter. *Breakthr. Stat. Methodol. Distrib.* **1992**, 492–518.
50. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
51. Zhu, X.; Lei, Z.; Yan, J.; Yi, D.; Li, S.Z. High-fifidelity pose and expression normalization for face recognition in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 787–796.
52. Fanelli, G.; Dantone, M.; Gall, J. Random forests for real time 3d face analysis. *Int. J. Comput. Vis.* **2013**, *101*, 437–458. [CrossRef]