This ACCV 2020 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ACCV 2020
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

Rotation Axis Focused Attention Network (RAFA-Net) for Estimating Head Pose

Ardhendu Behera[0000−0003−0276−9000], Zachary Wharton[0000−0001−5983−6468], Pradeep Hewage[0000−0002−6909−3546], and Swagat Kumar[0000−0001−7405−3445]

Computer Science, Edge Hill University, Ormskirk Lancashire L39 4QP, UK
{beheraa, zachary.wharton, pradeep.hewage, kumars}@edgehill.ac.uk

Abstract. Head pose is a vital indicator of human attention and behavior. Therefore, automatic estimation of head pose from images is key to many applications. In this paper, we propose a novel approach for head pose estimation from a single RGB image. Many existing approaches often predict head poses by localizing facial landmarks and then solve 2D to 3D correspondence problem with a mean head model. Such approaches rely entirely on the landmark detection accuracy, an ad-hoc alignment step, and the extraneous head model. To address this drawback, we present an end-to-end deep network, which explores rotation axis (yaw, pitch and roll) focused innovative attention mechanism to capture the subtle changes in images. The mechanism uses attentional spatial pooling from a self-attention layer and learns the importance over fine-grained to coarse spatial structures and combine them to capture rich semantic information concerning a given rotation axis. The evaluation of our approach using three benchmark datasets is very competitive to state-of-the-arts, including with and without landmark-based methods. Code can be found at https://github.com/ArdhenduBehera/RAFA-Net.

1 Introduction

Head pose estimation aims to infer the orientation of a person's head relative to the camera view. It is often represented using a 3D vector containing the Euler angles of yaw, pitch and roll. It is a key to many real-world applications such as aiding eye gaze estimation, human attention modeling, driver behavior understanding, human-robot social interactions, face alignments, human-computer interactions and many more. Over the past 20 years, there is a significant advancement in face detection. However, the reliable estimation of head poses from a single RGB image is still challenging, particularly in unconstrained 'in the wild' scenarios. For extreme poses, even face detection is arguably still difficult.

Estimating head pose from an image is essentially solving the mapping problem between 2D and 3D spaces. Traditionally, this is carried out using two steps: 1) detecting 2D facial landmarks in the target face, and 2) establishing the correspondence between landmarks and a head template [1–4]. The recent surge in deep Convolutional Neural Networks (CNNs) has significantly influenced in detecting and localizing facial landmarks [5–8]. This is mainly due to the fact

that deep models are often robust to extreme poses and occlusions, encouraging improvements in performance. Most of these models are aimed to estimate head poses and detect facial landmark, jointly. Moreover, the main goal is to improve the accuracy of facial landmark detection with the help of head poses, resulting in head pose estimation itself is not accurate enough.

There is no doubt that the advancement of deep CNNs has significantly improved landmarks detection accuracy. However, there are still possibilities of introducing errors in landmark-based head pose estimation. These are: 1) an insufficient number of detected landmarks, 2) quality of the head models/templates, and 3) their adaptation to each individual is also influenced by the model deformation, which is computationally expensive. To address this, there is a significant interest in estimating head poses directly from image intensities [9–14, 7, 6]. Existing works also use multimodal information such as RGB+depth images [15–18] and temporal knowledge from videos [19, 20] to improve the head pose estimation accuracy. It has significantly helped in improving performance but has its drawbacks. For example, depth cameras to capture depth information can be challenging to use in outdoors, and uncontrolled environments and are not always available. Therefore, there is a need for fast and reliable monocular image-based head pose estimation. On the other hand, temporal information in videos involving detection and tracking of heads could guide the pose estimation. However, modeling temporal information is often achieved with the use of recurrent networks, which are usually computationally expensive.

Our contribution: We propose a landmark-free end-to-end regression model called RAFA-Net (Rotation Axis Focused Attentional Network) for head pose estimation from monocular images. Head poses in monocular images often exhibit subtle changes. Deep models over the full images with distinctive classes have shown great success, but it raises the question about their performance in recognizing fine-grained changes [21]. Therefore, there is a need for learning meaningful features linking fine-grained changes for performing regression. One way to address this problem by adapting statistical pooling or aggregation approaches [22, 23], which learn high-level representative features from the low-level local features. However, such approaches often do not consider the spatial relationships, resulting in them being unable to capture the spatial structure, which is necessary for modeling fine-grained changes. Thus, we propose a novel attentional spatial pooling that *learns to distill* fine-grained to coarse spatial structures and combines them based on their importance to capture rich semantic information for estimating head poses. Moreover, the pooling module is attached to a given rotation axis (yaw, pitch and roll) to capture specific fine-grained changes in the image intensities for accurate head pose estimation.

Our attentional pooling can be interpreted as a more flexible and versatile pooling tool. Conventional pooling uses a pre-defined fixed window size (RoI), strides and types for a given task. Whereas, in our case, we pool features from a set of possible pooling (a combination of types, size and strides) covering smaller area to wider area with a more versatile approach to capture both local and global structures. Our approach is very similar to the recent work of deformable

RoI pooling [24, 25] for object detection and semantic segmentation. However, to generate new feature maps, we use attentional RoI pooling, which learns to distill the intrinsic consistency between informativeness of pooled features and their usefulness in estimating poses. Moreover, our attention map conveys how much to concentrate a given RoI in focus conditioned on all other RoIs. Whereas, deformable RoI pooling generates feature map via weighted summation of RoIs.

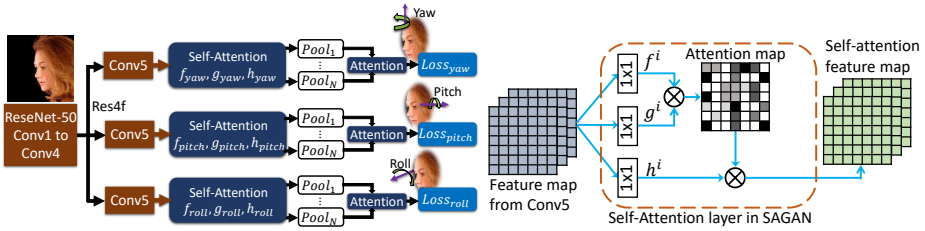
2 Related Work

Facial landmark-based approaches: Detected 2D facial keypoints are used to estimate head poses using 3D techniques such as POSIT [26]. The face alignment is often carried out using regression [1, 3, 4, 27], as well as model-based approaches [28–30]. Lately, CNN models [2, 31] for estimating 3D faces have shown superior performance. However, these approaches require manually annotated ground-truth, which is laborious, time-consuming, and often experts cannot accurately assign landmark locations in low-resolution images.

Landmark-free approaches: To address the above drawback, recently, there is a significant interest in estimating head poses directly from the image intensities using deep networks [9–14]. Such approaches often encounter problems due to illumination variations or poor illumination during night time. To overcome this, researchers have explored the complementary depth information for higher accuracy [16–18, 32]. Similarly, sequential knowledge from videos is explored in [19, 20] to benefit from the temporal coherence by using particle filters and recurrent networks to track facial features over time for improved head pose estimation.

Multi-tasking approaches: Facial modeling and analysis are multi-task learning (e.g. face detection, person identification, landmark detection, recognizing emotions, etc.) and is closely linked to head poses. Therefore, it has been shown that learning-related tasks jointly achieve better performance than individually [4–8]. Most of these methods are based on end-to-end deep learning models.

Attention-based approaches: Attention mechanism in machine learning is influenced by the human perception that focuses on selective parts of image/video to acquire salient information at specific locations and times. It has drawn increasing interest in solving machine translation [33, 34], image/video captioning [35–37], image/video recognition [38–40] and visual question answering [41] problems. Head pose estimation using attention mechanism is yet to be explored. This could be due to the head pose is a regression model, whereas, most of the existing approaches are applied to the classification of sequence mapping. Recently, Yang et al. [11] use a spatial attention proposal for refining regression values to estimate head poses. Our method is different from them since we use attentional spatial pooling and *learn to attend* the importance of fine-grained to coarse spatial structures to capture the subtle changes in images. Moreover, our model learns the rotation-axis specific subtle changes for estimating head poses. Our method is simple yet efficient and can be easily applicable to other applications.



(a) Proposed RAFA-Net based on ResNet (b) Rotation axis focused self-attention

Fig. 1: RAFA-Net for estimating head poses by introducing rotation axis-specific (yaw, pitch and roll) self-attention and attentional pooling components.

We revisit many of the above approaches (especially landmark-free methods, residual networks and attention mechanism) for advancing knowledge and solving the head pose estimation problem. We benefit from the well-known and very efficient ResNet architecture with simple yet efficient network modifications to capture salient information linking fine-grained changes in monocular images for estimating the head poses. We emphasize that our contributions include not only the modification to ResNet architecture but also an empirical study on the role of attentional spatial pooling in improving pose estimation accuracy.

3 Proposed Approach (RAFA-Net)

RAFA-Net is based on the ResNet model [42], which is adapted by introducing rotation axis-specific self-attention and attentional pooling layer to estimate head orientation represented using yaw, pitch and roll (Fig. 1a). In a CNN, initial layers learn more generic features (e.g. edges, corners, color blobs, etc.). As we move towards the output layer, the network gradually moves from generic to task-specific high-level features (e.g. structural/shape information). We explore this by modifying the last convolution (`Conv5`) layer of the ResNet-50 to learn rotation axis-specific high-level structural information. As a result, we use three parallel `Conv5` layers focusing on the respective yaw, pitch and roll axes. The output (width W, height H and C channels) of axis-specific `Conv5` is used to compute the respective bandwidth-specific self-attention map $\alpha = \{\alpha^{yaw}, \alpha^{pitch}, \alpha^{roll}\}$ (Fig. 1b) to capture important cues focusing on spatial changes. For each axis, the goal is to explicitly learn the relationships between features (dimension C) spatially located in a given resolution of $W \times H$. It conveys how much to focus the features at a given spatial location when synthesizing feature in another location. To achieve this, we compute the self-attention map $(\alpha^{yaw}, \alpha^{pitch}, \alpha^{roll})$ by adapting the SAGAN concept [43] in which the *query*, the *key*, and the *value* are all the same. For clarity, we describe the process for computing α and is the same for each rotation axis. Let's consider $x \in \mathbb{R}^{W \times H \times C}$ is the output of a `Conv5` layer for an image I (Fig. 1a). To compute α, the feature x is first transformed into the concept of *key* $\boldsymbol{f}(x) = \boldsymbol{W_f}x$, *query* $\boldsymbol{g}(x) = \boldsymbol{W_g}x$, and *value*

$\boldsymbol{h}(x) = \boldsymbol{W_h}x$. The element $\alpha_{i,j}$ indicates the extent to which the α attends to the j^{th} location while focusing on the i^{th} position in x and is computed using softmax function. It is used to compute the output o_j, which is a column vector of final output $o = (o_1, o_2, \ldots, o_j, \ldots, o_{W\times H}) \in \mathbb{R}^{W\times H\times C}$ and is computed as:

$$o_j = \sum_{i=1}^{W\times H} \alpha_{i,j}\boldsymbol{h}(x_i), \text{where } \alpha_{i,j} = \frac{\exp(s_{i,j})}{\sum_{j=1}^{W\times H}\exp(s_{i,j})}, s_{i,j} = \boldsymbol{f}(x_i)^T\boldsymbol{g}(x_j) \quad (1)$$

$\boldsymbol{W_f}$, $\boldsymbol{W_g}$ and $\boldsymbol{W_h}$ are all 1×1 convolution filters. We compute $o = \{o^k\}$ for each axis $k \in \{yaw, pitch, roll\}$. In addition, we also learn axis-specific scalar β^k and multiply with the output o^k and then add it with the input feature map x^k.

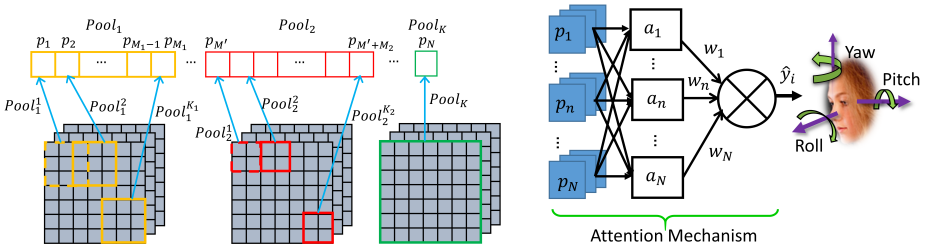
$$\hat{x}^k = \beta^k o^k + x^k, \text{ where } k \in \{yaw, pitch, roll\} \quad (2)$$

β^k is initialized to zero. It allows the network to first rely on the axis-specific local cues and then gradually learns to assign more weight to the global evidence. Afterwards, $\hat{x}^k$ is passed to our novel attentional spatial pooling (Fig. 2a).

3.1 Attentional Spatial Pooling

Spatial pooling is a standard building block of modern CNNs. In terms of the receptive field, there are two types (*local* or *global*) of spatial pooling widely used. Usually, *global* pooling often substitutes for the `FC` layer in many CNNs [42, 44, 45] via spatially squeezing the feature map tensor into a vector of channel dimensionality and is fed into the final classification/regression layer. However, global pooling loses the spatial structure and therefore, it might not be able to capture the subtle changes in images containing face orientations. In contrast, *local* pooling in CNNs [45–48] is commonly used to reduce spatial resolution with increasing robustness in variation (e.g. translation) against input images. It deals with only local features in the receptive field [49, 50]. To get the best out of both, one needs an appropriate method to combine them.

To address this, we propose a novel trainable *hybrid* spatial pooling, which employs attention mechanism and learns importance over fine-grained (local) to coarse (global) structures and combines them to attain rich semantic information in images. Our approach learns rotation axis specific pooling module (see Fig. 1a) that combines a various combination of pooling parameters (sizes, types and strides) and adaptively tunes it without manually fixing it beforehand. Let's consider our pooling module $\mathcal{P} = \{pool_1, pool_2, \ldots, pool_K\}$ consists of K possible poolings (i.e. combination of types, size and strides). One module per axis $i \in \{yaw, pitch, roll\}$ receives input from the respective self-attention map $\hat{x}^i \in \mathbb{R}^{W\times H\times C}$. Each $pool_k$ $(k = 1 \ldots K)$ is a unique combination of pooling size, type and stride resulting in a fixed number of RoI (region-of-interest) per $pool_k$ to be used for pooling (i.e. spatial sliding positions) in the spatial resolution of $W \times H$ (Fig. 2a). As a result, there is K_1 RoIs in $pool_1 = \{pool_1^1, pool_1^2, \ldots, pool_1^{K_1}\}$, K_2 RoIs in $pool_2 = \{pool_2^1, pool_2^2, \ldots, pool_2^{K_2}\}$



(a) Attentional spatial pooling (b) Weighting each pooled feature

Fig. 2: Before computing the loss for the respective Euler angle, we use a novel attentional spatial pooling from self-attention layer to capture rich semantic information representing the given rotation axis-specific angle. Afterwards, we combine them with our innovative attention mechanism.

and so on, until a single RoI in $pool_K = \{pool_K^1\}$ since $pool_K$ is the *global* pooling considering whole spatial resolution of $W \times H$. We concatenate all RoIs i.e. $\mathcal{P} = (pool_1^1, \ldots, pool_1^{K_1}, pool_2^1, \ldots, pool_2^{K_2}, \ldots, pool_K)$ over K pooling combinations and represent as $\mathcal{P} = (p_1, p_2, \ldots, p_N)$, where N is the total number of RoIs (see Fig. 2a). Each element p_n $(n = 1 \ldots N)$ is a RoI-pooled feature. During decision making, our pooling module $\mathcal{P}$ *learns to focus* on each p_n by its importance. We achieve this by introducing an attention-focused learnable parameter θ_a to compute high-level feature encoding $\mathbf{x} = f_a(p_n, a_n; \theta_a)$, where a_n is the attention-focused representation of RoI-pooled feature p_n and f_a is a mapping function. The element a_n is computed using the weighted summation of all other RoI-pooled features $p_{n'}$ and their similarity (measured in the form of probability) $\tau_{n,n'}$ to a given feature p_n in focus. This novel attention mechanism is implemented using an LSTM (Long Short-Term Memory) cell as follows:

$$a_n = \sum_{n'=1}^{N} \tau_{n,n'} p_{n'}, \text{ where } \tau_{n,n'} = \frac{\exp(\sigma_{n,n'})}{\sum_{n'=1}^{N} \exp(\sigma_{n,n'})}, \tag{3}$$
$$\sigma_{n,n'} = W_\sigma \rho_{n,n'} + b_\sigma, \text{ and } \rho_{n,n'} = tanh(W_\rho p_n + W_{\rho'} p_{n'} + b_\rho)$$

W_ρ and $W_{\rho'}$ are weights matrices for the respective RoI pooling combinations n and n'; W_σ is their non-linear fusion. $\tau_{n,n'}$ is computed from $\rho_{n,n'}$ using the sigmoid function; b_ρ and b_σ are the biases. The attention-focused representation a_n conveys how much to *attend the RoI-pooled feature* p_n in focus *conditioned on all other RoI-pooled features* (Fig. 2b). Finally, high-level feature map $\mathbf{x}$ for a given axis (yaw, pitch and roll) is computed by a weighted summation of all the pooling combinations using the attention importance weight w_n.

$$\mathbf{x} = \sum_{n=1}^{N} a_n w_n, \text{where } w_n = \frac{\exp(\psi_n)}{\sum_{j=1}^{N} \exp(\psi_j)} \text{ and } \psi_n = W_\psi a_n + b_\psi \tag{4}$$

The weight matrix W_ψ and bias b_ψ are learned. The attention importance score w_n for each a_n is constructed via probability distribution over the pooling representations using the sigmoid function. This approach is similar to the attention-based approach used to solve machine translation problems [51] in which the model automatically searches for parts of a source sentence that are relevant to predicting a target word. The difference is that we do not consider the sequential information. The final feature map $\mathbf{x}$ is used as an input to a final linear regression layer to solve the head pose estimation. Our attentional spatial pooling module consists of learnable parameter $\theta_a = \{W_\rho, W_{\rho'}, W_\sigma, W_\psi, b_\rho, b_\sigma, b_\psi\}$ for each rotation axis (yaw, pitch, and roll) o estimate axis-specific pose angles.

3.2 Learning

RAFA-Net is trained in an end-to-end fashion with the default ResNet input image size of 224×224. The model takes a set of training images $I = \{I^m | m = 1, \ldots, M\}$ and the respective head pose value of yaw (y^m_{yaw}), pitch (y^m_{pitch}) and roll (y^m_{roll}) in Euler angle (radian). The aim is to train the model to predict $\hat{y}^m_{yaw}$, $\hat{y}^m_{pitch}$, $\hat{y}^m_{roll} = model(I^m)$ for a given image I^m by minimizing combined regression loss (L_{MSE}), which is computed as a Mean Squared Error.

$$L_{MSE} = \frac{1}{M}\sum_{m=1}^{M} \underbrace{(y^m_{yaw} - \hat{y}^m_{yaw})^2}_{\text{Yaw MSE Loss}} + \underbrace{(y^m_{pitch} - \hat{y}^m_{pitch})^2}_{\text{Pitch MSE Loss}} + \underbrace{(y^m_{roll} - \hat{y}^m_{roll})^2}_{\text{Roll MSE Loss}} \quad (5)$$

4 Experiments

4.1 Implementation

RAFA-Net is implemented using Keras with TensorFlow as a backend. The convolutional layers (Conv1 to Conv5) are pre-trained layers from the ResNet-50 model [42] trained on the ImageNet [52] dataset. The model is trained with 150 epochs (32 batch size) using RMSProp optimizer [53] with a learning rate of 0.001 and rho of 0.9. The experiments are performed on a Linux PC (Ubuntu OS, Intel Core i9 9820X) with an NVIDIA Titan V GPU (12GB).

For an input image of size $224 \times 224 \times 3$, the self-attention module's output feature map resolution is $7 \times 7 \times 2048$ (Fig. 1b). For our attentional spatial pooling module $\mathcal{P}$, we experimentally found that *max pooling* is the best possible pooling type for this task. Given the spatial resolution of 7×7, we use pooling sizes of 2, 3, 4, 5 and 7. Similarly, we use the pooling stride of 2 and 3.

4.2 Datasets and Evaluation Strategies

There are a number of datasets produced so far for head pose estimation [54, 55]. Often facial landmarks are used to generate the ground-truth head poses by fitting a mean 3D face with the POSIT algorithm [26] since it is difficult to

(a) 300W-LP (b) AFLW2000 (c) BIWI

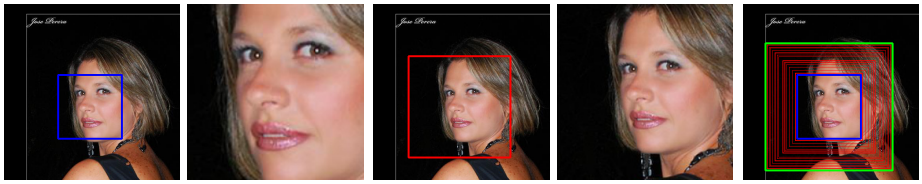
Fig. 3: Example images from three datasets: a) 300W-LP synthetic [31] - the various rendered head poses. b) AFLW2000 [31] - head poses from real-world images with varying background and lighting conditions. c) BIWI [17] - head poses from RGB-D images collected under a controlled environment.

precisely measure (or manually annotate) them. This approach works well for smaller angle head poses. However, it does not work well for large head poses due to the accuracy of facial landmark detection deteriorates in large poses and is mainly due to occlusion. For our experiments, we have used the three most popular datasets: 1) 300W-LP [31], 2) AFLW2000 [31], and 3) BIWI [17]. A few examples from these datasets are shown in Fig. 3. The 300W-LP [31] dataset is derived from the 300W dataset [55], which is a collection of several datasets for face alignment with 68 facial landmarks. It uses face profiling with 3D image meshing to generate 61,225 images of faces having large poses and further expanded to 122,450 faces with flipping. It is called as the 300W across Large Poses (300W-LP) and is synthetically generated by predicting the depth of each face, and then its profile views are computed with 3D rotation. The AFLW2000 dataset [31] is the subset (first 2000 images) of the AFLW dataset [56], and consists of head pose with large variations, facial expressions, different illumination, and occlusion conditions. It provides ground-truth annotations consisting of 3D faces and the corresponding 68 3D landmarks. The BIWI dataset [17] contains 15,678 frames from 24 RGB-D videos of 20 subjects captured using a Kinect device. These videos are captured in a controlled environment, and the 3D model is fitted to the RGB-D videos to obtain the ground-truth head poses. The head poses angle ranges are $\pm 77^o$ for yaw, $\pm 60^o$ for pitch, and $\pm 50^o$ for the roll.

To compare the performance of the RAFA-Net with state-of-the-arts, we follow the standard evaluation strategies, which are: 1) train on the synthetic 300W-LP large dataset and test on the other two relatively small datasets (AFLW2000 and BIWI). 2) train the model using 70% of videos (16 videos) in the BIWI dataset and evaluate the rest 30% (8 videos). In all three datasets, we use the detected face bounding box provided by Shao et al. [13]. The standard evaluation metric of mean absolute error (MAE) is used. For each pose angle, the average prediction error in degrees over testing images is used for the comparison. We have also compared the average prediction error over three (yaw, pitch and roll) Euler angles to show the overall performance of the proposed approach.

4.3 Data Augmentation

We propose a novel data augmentation approach (Fig 4) and is inspired by the experiment carried out by Shao et al. [13] to measure the accuracy of their



(a) (b) (c)

Fig. 4: Data augmentation involving randomization of bounding box margin using a control parameter γ: a) original bounding box from a face detector ($\gamma = 0$) and the corresponding cropped and resized (224×224) image, b) bounding box with $\gamma = 0.3$ and the corresponding cropped and resized image, c) 20 randomly generated bounding boxes (red) between blue and green bounding box, i.e. $0 \leq \gamma \leq 0.5$. Best view in color.

model by selecting a different size of the bounding box (prior to training and evaluation) enclosing a face. Our is different since we randomly generate these bounding boxes during training using a control parameter γ. Let (b_x, b_y) is the provided top-left location of a square bounding box b with size b_s. The corresponding bottom-right corner location will be at $(b_x + b_s, b_y + b_s)$. The aim is to generate different locations of top-left $(b_x - \gamma b_s, b_y - \gamma b_s)$ and bottom-right $(b_x + b_s + \gamma b_s, b_y + b_s + \gamma b_s)$ corners using γ to control bounding box margins. We experimentally found that this randomization gives better generalization resulting in improved performance rather than using standard augmentation techniques such as random scaling, width and/or height sifting and cropping. For all our experiments, we have used $0 \leq \gamma \leq 0.5$.

4.4 Comparison with the State-of-the-Art (SotA) Methods

We first compare the SotA pose estimation methods trained on the 300W-LP [31] and tested on the AFLW2000 [31] and BIWI [17], respectively. The performance comparison is presented in Table 1. In this experiment, the training and testing datasets are very different. For example, 300W-LP is a synthetic one, while the BIWI and AFLW2000 consist of real images. The deep learning-based landmark-free approaches such as Hopenet [10], SSR-Net-MD [12], ResNet-BBM [13], FSA-Net [11] and our RAFA-Net perform better than the landmark-based ones (Dlib [1], 3DDFA [31], FAN [2], KEPLER [5] and Two-stage [3]) tested on both the BIWI and AFLW2000 datasets. This is mainly since the landmark-free approaches can better accommodate the domain discrepancies between training and testing datasets.

Train on 300W-LP and test on AFLW2000: Our RAFA-Net is significantly outperformed the SotA approaches (Table 1). Among the existing landmark-free approaches, FSA-Caps-Fusion [11] provides the best performance. It uses the capsule network [23] for feature aggregation and gives equal importance to

Table 1: Comparison with the state-of-the-art approaches, which are trained on 300W-LP [31] dataset and evaluated on the respective AFLW2000 [31] and BIWI [17] datasets. The average error is in Euler angles (degrees).

Method	**AFLW2000 dataset [31]**				**BIWI dataset [17]**			
	Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
Dlib (68 landmarks) [1]	23.15	13.63	10.55	15.78	16.76	13.80	6.19	12.25
3DDFA [31]	5.40	8.53	8.25	7.39	36.18	12.25	8:78	19.07
FAN (12 landmarks) [2]	6.36	12.28	8.71	9.12	8.53	7.48	7.63	7.88
KEPLER [5]	-	-	-	-	8.80	17.3	16.2	13.9
Two-stage [3]	11.92	8.25	7.47	9:21	9.49	11:34	6.00	8.94
Ground-truth landmarks [10]	5.92	11.76	8.27	8.65	-	-	-	-
Hopenet ($\alpha = 2$) [10]	6.47	6.56	5.44	6.16	5.17	6.98	3.39	5.18
SSR-Net-MD [12]	5.14	7.09	5.89	6.01	4.49	6.31	3.61	4.65
ResNet-BBM (K=0.5) [13]	5.07	6.37	4.99	5.48	4.59	7.25	6.15	6.00
FSA-Caps-Fusion [11]	4.50	6.08	4.64	5.07	**4.27**	**4.96**	**2.76**	**4.00**
RAFA-Net(Ours: $\gamma = 0.3$)	**3.60**	**4.92**	**3.88**	**4.13**	5.71	6.28	3.64	5.21
RAFA-Net(Ours: $\gamma = 0.2$)	**3.52**	**4.93**	**3.91**	**4.12**	5.67	6.26	3.60	5.17

(a) Yaw (b) Pitch (c) Roll

Fig. 5: Visualization of the proposed attentional spatial pooling using rotation axis specific class activation map. RAFA-Net is trained on the 300W-LP dataset and tested on AFLW2000. It shows the rotation axis specific representative features are used for angle estimation. Visualization using BIWI images is included in the supplementary document.

different fine-grained feature mapping. We found that by providing weighted importance to the fine-grained to coarse spatial structures using our rotation axis-specific attentional spatial pooling produces more robust results. Similarly, KEPLER [5] aims to establish structural relationships between facial landmarks. Our approach is more effective than their iterative method since we learn the importance of fine-grained to coarse spatial structures and combine them by considering their importance to capture the rich semantic information. We have also evaluated the proposed approach by varying the bounding box parameter $0 \leq \gamma \leq 0.5$ during testing. We have found that the overall prediction error (MAE: 4.12) of our approach is the best (lower the better) for $\gamma = 0.2$, but the individual prediction error for pitch and roll is slightly better for $\gamma = 0.3$.

Train on 300W-LP and test on BIWI: The overall performance (MAE) of our RAFA-Net is inferior to the FSA-Caps-Fusion [11] and SSR-Net-MD [12] landmark-free approaches (Table 1). However, it is better than the ResNet-BBM

Table 2: Comparison with the state-of-the-art approaches using BIWI dataset [17]. There are three different evaluation methods (RGB only, RGB+Depth, and RGB+Time) by considering different modalities. The training (16 videos) data consists of 70% of the total videos (24) and the rest eight videos being used for testing. The average error is in Euler angles (degrees).

Method	Yaw	Pitch	Roll	MAE
RGB and Depth (RGB-D)				
DeepHeadPose [15]	5.32	4.76	-	-
Martin et al. [16]	3.60	2.50	2.60	2.90
3DMM [32]	**2.50**	**1.50**	**2.20**	**2.07**
RGB and Time				
VGG16+RNN [19]	3.14	3.48	2.60	3.07
Single RGB frame				
DeepHeadPose [15]	5.67	5.18	-	-
VGG16 [19]	3.91	4.03	3.03	3.66
SSR-Net-MD [12]	4.24	4.35	4.19	4.26
FSA-Caps-Fusion [11]	**2.89**	**4.29**	3.60	3.60
RAFA-Net (**ours**: $\gamma = 0.2$)	3.08	4.35	**2.85**	**3.43**
RAFA-Net (**ours**: $\gamma = 0.1$)	3.07	4.30	**2.82**	**3.40**

[13] and Hopenet [10]. Moreover, the estimated average error in pitch is better (6.26) than the landmark-free approaches except for the FSA-Caps-Fusion [11] (4.96). This could be due to the BIWI dataset is captured in a controlled environment with limited pose variations (yaw: $\pm 77^o$, pitch: $\pm 60^o$ and roll: $\pm 50^o$) and RGB-D videos are used to obtain the ground-truth head poses. Whereas, AFLW2000 consists of head poses with large variations ($\pm 99^o$) and is consistent with the training dataset 300W-LP. Nevertheless, our RAFA-Net performs significantly better than the landmark-based ones (Dlib [1], 3DDFA [31], FAN [2], KEPLER [5] and Two-stage [3]) on this dataset.

Train and test on BIWI: In this experiment, we compare the performance using only the BIWI dataset. We use the same train and test split, as provided in [11]. The dataset consists of RGB-D sequences, including color and depth information. The overall performance (MAE) of our approach is better than all other methods in its peer group (single RGB frame only). For individual yaw and pitch Euler angles, our method is very close to the FSA-Caps-Net [11] (e.g. Yaw: 2.89 vs 3.07 and Pitch: 4.29 vs 4.30). For roll, our approach is significantly better than the existing approaches. We also report the existing approaches, which combine the different modalities (RGB+Depth and RGB+Time) for improving performance. Our approach does not perform equally well compared to these methods, which combine multimodal data, but not too far from them. Additionally, our RAFA-Net performs significantly better than the multimodal approaches [19], [15], [16] in estimating yaw. Similarly, for pitch, our RAFA-Net is better than the DeepHeadPose [15] that uses RGB and depth information.

Table 3: Ablation study involving the performance of individual components. Performance comparison using our novel bounding box (BB) augmentation versus standard spatial augmentation (random scaling, shifting and cropping), as well as rotation axis-specific attentional pooling versus single attentional pooing. The respective model is trained on 300W-LP [31] dataset and evaluated on the respective AFLW2000 [31] and BIWI [17] datasets.

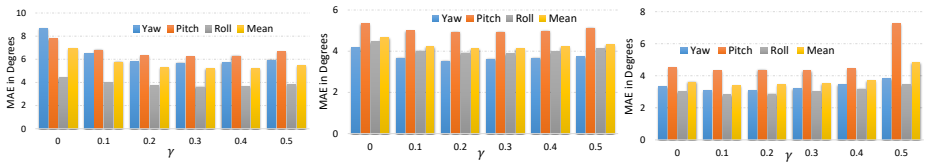
Experiment using	**AFLW2000 [31]**				**BIWI [17]**			
	Yaw	**Pitch**	**Roll**	**MAE**	**Yaw**	**Pitch**	**Roll**	**MAE**
Self-attn. only ($\gamma = 0.3$)	4.19	5.83	4.44	4.82	6.26	9.50	4.36	6.71
Attn. pooling only ($\gamma = 0.3$)	3.34	5.61	3.91	4.29	5.16	9.93	3.56	6.22
No BB augmentation	4.22	5.87	4.53	4.53	7.46	10.31	4.01	7.26
RAFA-Net (single attn)	3.93	5.51	4.13	4.52	6.85	8.45	4.18	6.49
RAFA-Net ($\gamma = 0.2$)	3.52	4.93	3.91	4.12	5.67	6.26	3.60	5.17
	Evaluation using spatial augmentation only							
Self-attn. only	5.83	6.54	5.74	6.04	9.00	7.42	5.67	7.36
Attn. pooling only	4.35	5.89	4.71	4.99	7.32	7.98	4.90	6.73
RAFA-Net	3.63	5.55	3.57	4.25	5.56	6.02	4.54	5.37

Table 4: Ablation study involving the performance of individual components using our novel bounding box (BB) augmentation versus standard spatial augmentation (random scaling, shifting and cropping). RAFA-Net is trained and tested using BIWI [17] datasets.

Experiment	**BB Augmentation**				**Spatial Augmentation**			
BIWI [17]	**Yaw**	**Pitch**	**Roll**	**MAE**	**Yaw**	**Pitch**	**Roll**	**MAE**
Self-attn. only ($\gamma = 0.4$)	3.84	5.26	3.91	4.34	4.62	5.99	4.67	5.10
Attn. pooling only ($\gamma = 0.3$)	4.23	6.00	4.11	4.80	4.98	5.54	4.25	4.92
No BB augmentation	4.37	4.74	3.98	4.36	-	-	-	-
RAFA-Net ($\gamma = 0.1$)	3.07	4.30	2.82	3.40	3.50	4.83	3.28	3.87

4.5 Ablation Studies

We have conducted an ablation study to understand the impact of the proposed novel attentional spatial pooling, rotation axis-specific self-attention, and our data augmentation approach involving the randomization of bounding box margin. The results are shown in Table 3 and Table 4. The performance of RAFA-Net trained on 300W-LP and tested on the respective AFLW2000 and BIWI datasets are presented in Table 3. It is observed that the MAE of our attentional spatial pooling is better than the self-attention, as well as our model without a randomized bounding box. Moreover, for AFLW2000 dataset, the performance of each component in Table 3 is better than the previous best FSA-CAPS-Fusion (MAE: 5.07) [11]. This justifies the benefits of each component. Among the three components, our novel spatial attentional pooling module is the best performer. This proves the significance of the proposed spatial pooling. We have also compared the performance of RAFA-Net using our novel bounding



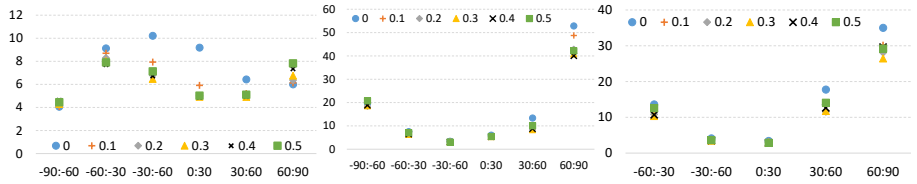
(a) Test AFLW2000 (b) Test BIWI (c) Train and test BIWI

Fig. 6: Effect of bounding box margin (control with parameter $0 \leq \gamma \leq 0.5$) on pose estimation error. RAFA-Net is trained on 300W-LP and evaluated on the respective a) AFLW2000 and b) BIWI datasets. c) RAFA-Net is trained and tested using BIWI dataset. Mean is the average of yaw, pitch and roll.

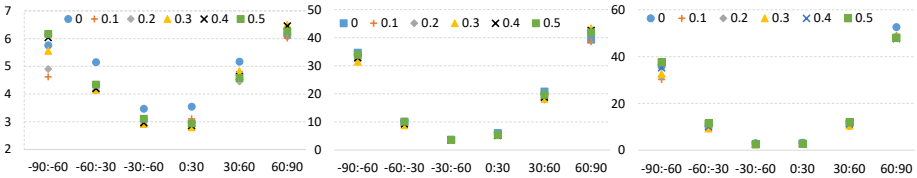
box augmentation versus standard spatial augmentation (random scaling, shifting and cropping). The proposed bounding box augmentation outperforms the standard spatial augmentation (Table 3). A similar trend is also observed when tested on the BIWI dataset. We have also assessed the performance of the above components using the BIWI dataset (train and test like in Table 2). The results are presented in Table 4. The rotation axis-specific attentional spatial pooling is also compared with the single attentional pooling predicting yaw, pitch and roll. The axis-specific attentional pooling is outperformed the single one (Table 3).

We have also carried out the ablation study for understanding the influence of bounding box margin parameter γ while testing. During training, we randomize the value of $0 \leq \gamma \leq 0.5$ while selecting the size of the bounding box enclosing a face (Fig. 4). During testing, we vary the value of γ from 0 to 0.5 with an increment of 0.1 and evaluate the prediction error. The results are reported in Fig 6. It shows the result of our model trained on 300W-LP [31] and evaluated on the respective AFLW2000 [31] and BIWI [17] datasets. One can observe that as the value of γ increases, the prediction error decreases (less the better) and reaches a minimum at $\gamma = 0.2$ and then increases. A similar trend is observed for $\gamma = 0.1$ when the model is evaluated using BIWI (Fig. 6c). Shao et al. [13] have also studied the effect of bounding box margin on prediction accuracy. However, our approach is different from them since we use the randomization of γ during training and evaluate the prediction accuracy with different γ values during testing. Whereas, they use the same fixed value during training and testing.

We have also studied the impact of bounding box margin parameter γ on different angle ranges (-90:-60, -60:-30, -30:0, 0:30, 30:60, and 60:90) using our RAFA-Net. The model is trained on 300W-LP [31] and tested on AFLW2000 [31] and BIWI [17] datasets. The results are presented in Fig. 7. It is evident that the estimation of the yaw angle is accurate for a wide range of angles, whereas the pitch and roll tend to be inaccurate for larger angles (absolute). This trend is observed in both BIWI and AFLW2000 datasets. It is also observed that the pitch and roll lean to insensitive to the γ values for smaller angles; however, they tend to be sensitive for larger angles. A noticeable observation is that yaw is sensitive to the γ values for a wide range of angles. Therefore, the optimal



(a) BIWI - Yaw (left), Pitch (middle) and Roll (right)



(b) AFLW2000 - Yaw (left), Pitch (middle) and Roll (right)

Fig. 7: Effect of bounding box margin (control with parameter $0 \leq \gamma \leq 0.5$) on average pose estimation error in degrees (y-axis) on different angle ranges (x-axis) for yaw, pitch, and roll. Our model is trained on 300W-LP and tested on the BIWI and AFLW2000 datasets.

value of γ has influenced the overall estimation accuracy. We have included the quantitative values in tabular form in the supplementary document.

5 Conclusion

In this paper, we have proposed a simple yet effective way to learn the importance of meaningful salient features in modeling fine-grained changes for head pose estimation using monocular images. By defining *learn to attend* weighting function via exploring attentional pooling mechanism, we are able to learn the importance of fine-grained to coarse spatial structures and combined them based on their importance to capture rich semantic information to solve the problem in hand. The proposed attentional pooling is employed to capture rotation axis specific semantic information, and our experiments have shown that the approach is better than the state-of-the-art methods. The proposed approach has demonstrated to improve the head pose estimation accuracy; however, we believe that this idea can be adapted to other regression and classification problems. Future work will be to apply the proposed technique for multi-task learning linking facial expression analysis, modeling, and recognition.

Acknowledgements

This research was supported in part by the UKIERI-DST (CHARM) under grant DST UKIERI-2018-19-10. The GPU used in this research is generously donated by the NVIDIA Corporation.

References

1. 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. (2014) 1867–1874
2. 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. (2017) 1021–1030
3. Lv, J., Shao, X., Xing, J., Cheng, C., Zhou, X.: A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 3317–3326
4. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE conference on computer vision and pattern recognition, IEEE (2012) 2879–2886
5. Kumar, A., Alavi, A., Chellappa, R.: Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. In: 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE (2017) 258–265
6. Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R.: An all-in-one convolutional neural network for face analysis. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE (2017) 17–24
7. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **41** (2017) 121–135
8. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: European conference on computer vision, Springer (2014) 94–108
9. Kuhnke, F., Ostermann, J.: Deep head pose estimation using synthetic images and partial adversarial domain adaption for continuous label spaces. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 10164–10173
10. 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. (2018) 2074–2083
11. Yang, T.Y., Chen, Y.T., Lin, Y.Y., Chuang, Y.Y.: Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 1087–1096
12. Yang, T.Y., Huang, Y.H., Lin, Y.Y., Hsiu, P.C., Chuang, Y.Y.: Ssr-net: A compact soft stagewise regression network for age estimation. In: IJCAI. Number 6 (2018) 7
13. Shao, M., Sun, Z., Ozay, M., Okatani, T.: Improving head pose estimation with a combined loss and bounding box margin adjustment. In: 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019). (2019) 1–5
14. Chang, F.J., Tuan Tran, A., Hassner, T., Masi, I., Nevatia, R., Medioni, G.: Faceposenet: Making a case for landmark-free face alignment. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2017) 1599–1608
15. Mukherjee, S.S., Robertson, N.M.: Deep head pose: Gaze-direction estimation in multimodal video. IEEE Transactions on Multimedia **17** (2015) 2094–2107

16. Martin, M., Van De Camp, F., Stiefelhagen, R.: Real time head model creation and head pose estimation on consumer depth cameras. In: 2nd International Conference on 3D Vision. Volume 1., IEEE (2014) 641–648
17. Fanelli, G., Dantone, M., Gall, J., Fossati, A., Van Gool, L.: Random forests for real time 3d face analysis. International journal of computer vision **101** (2013) 437–458
18. Meyer, G.P., Gupta, S., Frosio, I., Reddy, D., Kautz, J.: Robust model-based 3d head pose estimation. In: Proceedings of the IEEE international conference on computer vision. (2015) 3649–3657
19. Gu, J., Yang, X., De Mello, S., Kautz, J.: Dynamic facial analysis: From bayesian filtering to recurrent neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 1548–1557
20. Chrysos, G.G., Antonakos, E., Snape, P., Asthana, A., Zafeiriou, S.: A comprehensive performance evaluation of deformable face tracking "in-the-wild". International Journal of Computer Vision **126** (2018) 198–232
21. Girdhar, R., Ramanan, D.: Attentional pooling for action recognition. In: Advances in NIPS. (2017) 33–44
22. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 5297–5307
23. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. In: Advances in neural information processing systems. (2017) 3856–3866
24. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 9308–9316
25. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. (2017) 764–773
26. Dementhon, D.F., Davis, L.S.: Model-based object pose in 25 lines of code. International journal of computer vision **15** (1995) 123–141
27. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. International Journal of Computer Vision **107** (2014) 177–190
28. Matthews, I., Baker, S.: Active appearance models revisited. International journal of computer vision **60** (2004) 135–164
29. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Computer vision and image understanding **61** (1995) 38–59
30. Liang, L., Xiao, R., Wen, F., Sun, J.: Face alignment via component-based discriminative search. In: European conference on computer vision, Springer (2008) 72–85
31. 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. (2016) 146–155
32. Yu, Y., Mora, K.A.F., Odobez, J.M.: Robust and accurate 3d head pose estimation through 3dmm and online head model reconstruction. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Ieee (2017) 711–718
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. (2017) 5998–6008

34. Cinar, Y.G., Mirisaee, H., Goswami, P., Gaussier, E., Aït-Bachir, A., Strijov, V.: Position-based content attention for time series forecasting with sequence-to-sequence rnns. In: International Conference on Neural Information Processing, Springer (2017) 533–544
35. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 4634–4643
36. Herdade, S., Kappeler, A., Boakye, K., Soares, J.: Image captioning: Transforming objects into words. In: Advances in Neural Information Processing Systems. (2019) 11137–11147
37. Pei, W., Zhang, J., Wang, X., Ke, L., Shen, X., Tai, Y.W.: Memory-attended recurrent network for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 8347–8356
38. Girdhar, R., Carreira, J., Doersch, C., Zisserman, A.: Video action transformer network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 244–253
39. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: Thirty-first AAAI conference on artificial intelligence. (2017)
40. Sharma, S., Kiros, R., Salakhutdinov, R.: Action recognition using visual attention. arXiv preprint arXiv:1511.04119 (2015)
41. Li, X., Song, J., Gao, L., Liu, X., Huang, W., He, X., Gan, C.: Beyond rnns: Positional self-attention with co-attention for video question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 8658–8665
42. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE CVPR. (2016) 770–778
43. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318 (2018)
44. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI. (2017) 4278–4284
45. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: IEEE CVPR. (2016)
46. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012) 1097–1105
47. Zhai, S., Wu, H., Kumar, A., Cheng, Y., Lu, Y., Zhang, Z., Feris, R.: S3pool: Pooling with stochastic spatial sampling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 4970–4978
48. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
49. Lee, C.Y., Gallagher, P.W., Tu, Z.: Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree. In: Artificial intelligence and statistics. (2016) 464–472
50. Saeedan, F., Weber, N., Goesele, M., Roth, S.: Detail-preserving pooling in deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 9108–9116
51. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
52. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV **115** (2015) 211–252

53. Bengio, Y., CA, M.: Rmsprop and equilibrated adaptive learning rates for non-convex optimization. corr abs/1502.04390 (2015)
54. Behera, A., Gidney, A.G., Wharton, Z., Robinson, D., Quinn, K.: A CNN model for head pose recognition using wholes and regions. In: IEEE Int'l Conf. on Automatic Face & Gesture Recognition (FG). (2019)
55. 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. (2013) 397–403
56. Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: 2011 IEEE international conference on computer vision workshops (ICCV workshops), IEEE (2011) 2144–2151