Person Gender Classification on RGB-D Data with Self-Joint Attention

XIAOXIONG ZHANG 1, SAJID JAVED2, JORGE DIAS2, AND NAOUFEL WERGHI2 (Member, IEEE)

1Department of Robotics, Khalifa University, P.O. Box: 127788, Abu Dhabi, UAE (e-mail: 100049862@ku.ac.ae)
2Department of Electrical Engineering And Computer Science, Khalifa University, P.O. Box: 127788, Abu Dhabi, UAE (e-mail: sajid.javed@ku.ac.ae; jorge.dias@ku.ac.ae; naoufel.werghi@ku.ac.ae)

Corresponding author: Naoufel Werghi (e-mail: naoufel.werghi@ku.ac.ae).

This publication is based upon work supported by the Khalifa University of Science and Technology under Award No. RC1-2018-KUCARS.

ABSTRACT
Automatic gender classification has many potential applications including automatic annotation of images, video surveillance, security, and human-computer interaction. In the last decades, many research works focused on classifying gender using the cues from 2D images of the person’s frontal view. This limits their application in the real world. Also, classifying a person’s gender from a different view, pose and scale are still challenging problems. RGB-D images, containing color and depth images, have more immunity to this problem than 2D images. Recent approaches using RGB-D images explored different combinations of features descriptors from depth and color images to classify the gender. These methods used low-level features extracted from the depth and color separately, ignoring the inter-dependency between the two modalities. In this article, we proposed a deep learning-based approach using a self-joint attention mechanism for human gender classification on RGB-D images. The proposed attention mechanism is designed to encode the inter-dependent information between the depth and color images to enhance the feature discrimination power. We benchmarked our method on a challenging gender dataset that consists of different views, poses, and scales. The presented method outperforms the state-of-art methods with an improvement in the accuracy of 5.2%, 7.5%, and 8.7% on three different test data.

INDEX TERMS Gender classification, Attention, RGB-D Image

I. INTRODUCTION
Automatic gender classification has received great attention in pattern recognition and computer vision communities over a while. It has a variety of applications including image search, automatic annotation of images, demographic research, video surveillance and security, human-computer interaction, psychological analysis, and commercial profiling.

Various gender classification methods have been proposed employing different cues from the appearance information [1] [2] [3], e.g. face, body, or non-appearance information [4] [5], e.g. voice, iris, fingerprint, gait. Most of the works focused on the face as it contains reliable and distinctive information. Hence, researchers investigated different techniques to find discriminative representations of the face to ensure reliable predictions. For instance, the Local Binary Pattern (LBP) histogram [6] has been used to extract shape and texture features from the face. However, LBP-type features are limited in describing the local texture and cannot capture the global dominant texture. Other approaches used global feature descriptors like Histogram of Oriented Gradients (HOG) [7]. Combining the representative local features and the global features [8] is proposed to ensure a representation insensitive to the effects of illumination, translation, and rotation.

With the advent of deep learning, Convolutional Neural Networks (CNN) have also been introduced in gender classification [1]. The expressive power of CNN is frequently adopted to extract high-level semantic features, which significantly boost the models’ performance. Despite the above advancements, 2D image-based methods are still facing constraints and limitations. A clear look of the face is required for most face-based approaches. Further, with the presence of occlusion, facial pose dynamic change, and illumination change, most of the approaches still suffer from imperfections.

Encompassing both depth and appearance information, 3D
data, e.g. RGB-D image, is intrinsically more immune to the aforementioned nuisance factors. The depth information contains profound geometric information of the person in 3D. However, there is limited work using 3D data for gender classification and depth information is not yet fully exploited in this field. Some research works [9] only focused on extracting pattern features from the depth image, like gait. Extracting the same type of features separately from the two modalities [10] is another preferable way to use the RGB-D data for gender classification. But the inter-dependency between the RGB and depth is not encoded in these features which limits the potential performance of their models. 3D point cloud data is also explored for gender classification using the geometrical feature of the body [3]. However, these features are hand-crafted and their performance degrades in challenging scenarios.

In this work, we proposed a deep learning method exploiting the complementarity of the appearance and shape information in the RGB-D data for gender classification. Inspired by visual object tracking [11] and object detection [12] methods, we employed a CNN with an attention module to learn representative features from the RGB-D data. The attention module serves as a feature fusion mechanism to find the inter-dependency between the RGB image and depth image. This mechanism boosts the discriminative power of the feature representation while enhancing the capacity of gender recognition under different view-angle, pose changing, and occlusion.

The proposed model contains three key components: a feature extraction module, a feature fusion module, and a prediction head. The feature extraction module consists of two branches of pre-trained CNN to extract features from the two modalities separately. The self-joint attention module is then used to fuse the features from the two modalities by learning the self-joint attention. A binary prediction head, consisting of a fully connected layer, is used to predict the class of input images. Compared to competitive methods, our method is distinguished by the following contributions:

- It explores the deep neural networks to learn discriminative features from RGB-D images to distinguish men and women.
- It proposes a self-joint attention module to capture the inter-dependency between the RGB image and depth image. The self-joint attention module effectively fuses the self-attentions of the two modalities hence improving the feature representation.
- It achieves the best performance compared to the state-of-the-art methods on the challenging RGB-D gender dataset including occlusion, poses variation, and scales changing scenarios.

The remainder of this paper is organized as follows. Section II briefly reviews related work. Section III describes the proposed gender classification method. Section IV discusses the experimental results. Section V concludes the paper.

II. RELATED WORK

While most of the previous work focuses on 2D images, some research works explored 3D data for automatic gender classification. Ballihi et al. [13] proposed classifying a person’s gender by measuring the similarity of the geometric shape from the face. The geometric shape features, e.g. geodesic paths and distances between facial curves, are extracted from the 3D face image captured by a laser scanner. Then these features are compared to their counterparts in random male and female templates to derive the feature vector signature. Boutella et al. [10] compared the performance of Kinect depth and RGB image for different face analysis tasks including gender classification. Four texture and shape features, i.e. LBP, local phase quantization(LPQ), HOG, and binarized statistical image features (BSIF), are extracted from both the depth and RGB image separately. Two corresponding SVM classifiers are used for classification. The results, obtained with three different face datasets, show that the depth image has a close performance to the RGB image on gender classification. Azzakhninia and et al. [14] proposed using face parts to improve the classification. They extracted LBP features from the face parts and trained an individual classifier for each part. Then a decision fusion scheme is applied to fuse the classifiers’ results. However, the LBP is limited in describing the local texture and cannot capture faithfully the global dominant one.

Instead of focusing on the face, other research works looked for cues from other parts, such as the body [3], and skeleton [5]. Exhibiting more exposure than face, the whole body area has been thought of as an opportune alternative despite its high variability in terms of appearance, poses, and scale. Linder et al. [3] considered classifying the person’s gender from the body using 3D data from the RGB-D images. They propose a tessellation method that extracts the local geometry features from the point cloud of the human body. Global texture descriptors like the Histogram of oriented gradients (HOG) and histogram of oriented depths (HOD), computed on the RGB and D image respectively, are used added to integrate the information from the two modalities in [15].

Kwon et al. [5] proposed a gait feature called joint swing energy (JSE) extracted from a sequence of 3D positions of skeletal joints to discriminate the gender. The intuition behind the JSE feature is that women and men walk differently which can be visualized from the pelvis and hip motion.

Recent deep-learning methods are also used in this field. Socher et al. [16] trained convolutional-recursive neural networks (CRNNs) on the RGB image and depth image separately to learn high-level features. The learned features from the two modalities are then concatenated to make predictions. Similarly, Bo et al. [17] employed a multi-layer sparse coding network to learn features from the RGB and Depth images separately. The embedding outputs from each layer are used to form a dictionary to do the classification. Different from their work, we propose employing an attention module to learn and integrate the features from the two modalities.
Overall, most of the works dedicated to gender classification use features of the face to discriminate the gender of the person. While using RGB-D data, most approaches still focus on extracting discriminative hand-crafted features. In addition, many methods separate the feature extraction for the two modalities, and the inter-dependent information is ignored.

### III. METHODOLOGY

The overall architecture of the proposed method is shown in Figure 1. We will elaborate on the three key components: feature extraction module, feature fusion module, and prediction head in Section III-A and we will illustrate the process of the proposed self-joint attention module in Section III-B.

#### A. OVERALL ARCHITECTURE

**Feature Extraction:** We adopted a deep [18] convolution neural network (CNN) as the backbone to extract representative features from the RGB image, $I^r \in \mathbb{R}^{H_0 \times W_0 \times 3}$, and depth image, $I^d \in \mathbb{R}^{H_0 \times W_0 \times 3}$, respectively. Conforming to the input format of CNNs, $I^d$ is expanded into a three-channel image, $I^{d'} \in \mathbb{R}^{H_0 \times W_0 \times 3}$, by duplicating itself on the other channels. $I^r$ and $I^{d'}$ are reshaped to the sizes $H_0, W_0 = 256$. The pipeline of the whole system is shown in Figure 1. The backbone contains two branches of CNN for RGB image and depth image respectively. The two branches are not sharing the weights. Several convolution layers in gray color are frozen while the convolution layers in blue are not sharing the weights. Several convolution layers in yellow color are fine-tuned during the training. For the RGB branch, we only re-trained the top convolution layers to learn the task-specific high-level features. Because the CNN is pre-trained on a large amount of data, we deemed keeping more shallow layers frozen is appropriate to derive adequate feature representation. However, for the depth branch, we unfroze more convolution layers from the top to the bottom for better learning, as the CNN is not exposed to depth data in the pre-training. The extracted features are denoted as $X^r \in \mathbb{R}^{H \times W \times C}$ and $X^d \in \mathbb{R}^{H \times W \times C}$ for RGB and depth respectively.

**Feature Fusion:** The feature fusion module is meant to encode the complementarity between the RGB and the depth image, thus enhancing the feature representation. We proposed a self-joint attention module to fuse the information of the two modalities. The detail of this module will be elaborated in Section III-B. In the self-joint attention module, the extracted features $X^r$ and $X^d$ are used to learn their self-attention maps $f^r$ and $f^d$ individually at first. Then their self-attention maps are combined and augmented with the residual connection to the features $X^r$ to obtain the attention-enhanced features $Z$.

**Prediction Head:** The prediction head consists of two fully connected layers and an average pooling layer shown as FCN in Figure 1. The average pooling layer is used to pull out the feature vector, $Z_a \in \mathbb{R}^{1 \times C}$ based on each channel of $Z$. $Z_a$ is then flattened and forwarded into the fully connected layers. The fully connected layers, containing a 512 neurons layer with RELU activation function and a 2 neurons layer with a softmax activation function, predict the class of gender using $Z_a$.

#### B. SELF-JOINT ATTENTION

Self-attention has been widely adopted to enhance feature discriminative [12]. Inspired by this latest work and the model proposed in the Transformer architecture in [19], we proposed a self-joint attention module, depicted in Figure 2,
to effectively fuse the features from different modalities.

Given the embedding from the RGB image, \( X^r \in \mathbb{R}^{H \times W \times C} \), and depth image, \( X^d \in \mathbb{R}^{H \times W \times C} \), where \( H \), \( W \), and \( C \) are the height, width, and channel size respectively, the self-attention module first projects the \( X^r \) and \( X^d \) into three subspaces using a \( 1 \times 1 \) convolutional layers. These three subspaces are named as the query \( q \in \mathbb{R}^{H \times W \times C'} \), key \( k \in \mathbb{R}^{H \times W \times C'} \), and value \( v \in \mathbb{R}^{H \times W \times C'} \). \( C' \) is the dimension of subspaces. Note that, because the RGB and depth images are going through mostly the same process, we simplify the math expression in the rest of the paper by setting \( X = X^r \) or \( X^d \) unless specified. We computed the projection as:

\[
q(X) = XW_q, \quad k(X) = XW_k, \quad v(X) = XW_v \quad (1)
\]

where \( W_q, W_k, \) and \( W_v \in \mathbb{R}^{C \times C'} \) are the embedding weights. In the experiment, we set \( C' = 0.5C \). The \( q \) and \( k \) are then used to find the self-attention map, \( f \). First, the \( q \) and \( k \) are reshaped into \( \mathbb{R}^{HW \times C'} \) and then implemented the element wise multiplication to get the \( f \) as:

\[
f(q, k) = \frac{(q(X)k(X))^\top}{\sqrt{C'}} \quad (2)
\]

where \( f \in \mathbb{R}^{HW \times HW} \). \( \sqrt{C'} \) is empirically used to normalized \( f \) according to [19]. Each row of \( f_i \), of \( f \) represents the attention scores in the \( i \)th position. \( f(X^r) \) and \( f(X^d) \) simply contain the self-attention maps of their own modality.

Depth is particularly strong cue for object shape and has the potential of enhancing the distinct part of the person. Therefore, we added the self-attention from the two modalities to generate joint-attention map. Because the attention scores of the self-attention map have high variance, we applied column-wise softmax normalization on the joint-attention map as:

\[
f_n(q, k) = \text{softmax}_{col}(\lambda_1 f^r(q^r, k^r) + \lambda_2 f^d(q^d, k^d)) \quad (3)
\]

where \( \lambda_1 \) and \( \lambda_2 \) are the weight factors. They are chosen in a learning manner and we will elaborate on that in Section III-C. \( f_n \) is the normalized joint-attention map.

We can find the self-attention \( A \in \mathbb{R}^{H \times W \times C'} \) by multiplying the normalized joint-attention map, \( f_n \), and the value, \( v^r \), from the RGB embedding as in (4). \( v^r \) is reshaped to the \( HW \times C' \) before the multiplication. The output is reshaped back to \( H \times W \times C' \).

\[
A(q, k, v) = f_n(q, k)v^r \quad (4)
\]

According to the finding in the transformer paper [19], it is beneficial to learn the attention from multiple pairs of query, key, and value, which are different projections of \( X \). The attention learned from multiple pairs of projections can jointly attend to information from different subspaces at different positions. These attentions are then concatenated and projected back to the \( C' \) dimensional subspace, yielding the multi-head attention. We also adopted this multi-head attention into the self-joint attention module as depicted in the following:

\[
\text{MultiHead}(X^r, X^d) = \text{Concat}(\text{head}_1, ..., \text{head}_n)W_0
\]

where \( W_0 \in \mathbb{R}^{nC' \times C} \) is \( 1 \times 1 \) convolutional layer, \( n \) is the number of head. We chose \( n = 2 \) in the experiment.

Finally, a residual connection adds the multi-head attention to the RGB embedding \( X^r \), yielding the attended features as the output from the self-joint attention module:

\[
Z = \text{Multihead}(X^r, X^d) + X^r
\]

where \( Z \) is in \( \mathbb{R}^{H \times W \times C} \).

C. SELECTIVE FUSION

The self-joint attention module considers the weights of the attention from RGB and depth equally. However, practically the contribution does not present the same level of reliability. For example, depth data might be noisy and inaccurate compared to its RGB counterpart. Assigning different weights for different positions of attention is an effective way to suppress unwanted attention. To this end, we employed a learnable scheme, proposed in [12], to assign the weights for the attention of different modalities in the self-joint attention module as shown in Figure 2(b). The fusion weights are also learned using the embedding of the RGB image, \( X^r \), and depth image, \( X^d \). A convolution layer with a softmax activation is added to generate weight map \( \alpha \) from the \( X^r \) and \( X^d \):

\[
\alpha = \text{softmax}([X^r, X^d]W_\alpha) \quad (5)
\]

where \([ \quad ]\) denotes matrix concatenation and \( W_\alpha \) is a \( 1 \times 1 \) convolution layer. The weight map \( \alpha \in \mathbb{R}^{H \times W \times 2} \) can be split into two weight maps \( \alpha^r \) and \( \alpha^d \in \mathbb{R}^{H \times W \times 1} \). Each of the weight maps represents the reliability at all positions of the attention of the corresponding modality.

Then the selective self-joint attention can be obtained by updating (3) with a weighted sum of the two modalities’ attentions:

\[
A(q, k, v) = \text{softmax}_{col}(\alpha^r \odot f^r(q^r, k^r) + \alpha^d \odot f^d(q^d, k^d)) \times v^r \quad (6)
\]

where \( \odot \) is the channel-wise Hadamard product.

D. TRAINING LOSS

Our model is trained in an end-to-end fashion, where the training loss is a binary cross-entropy loss:

\[
\text{Loss} = -\frac{1}{N} \sum_{i=1}^{N} y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i) \quad (7)
\]

where \( y_i \in [0, 1] \) is the ground truth, \( \hat{y}_i \) is the prediction, and \( N \) is the number of samples.
IV. EXPERIMENTS AND RESULTS

A. DATASETS

We benchmarked our model using the RGB-D gender dataset [3]. This dataset contains four sets of challenging scenes that capture 115 instances (60 females and 64 males) in different postures, such as standing, walking, and interacting with the camera. In set-1, each person stands at a 2.5m distance from the camera and faces eight different directions. Set-2 and set-3 consist of a person walking in a circle path following clockwise and counter-clockwise directions respectively. Set-4 simulates human close-up interaction with the camera and the person goes back and forth in front of the camera. In this set, due to the limit of the camera’s view, most of the time only part of the persons’ body is visible. The dataset also includes some scenarios of people wearing backpacks, carrying luggage, and in wheelchairs. The dataset is recorded using a Kinect-2 camera, in which depth has a range from 0.5m to 4.5m. The camera keeps static during the recording.

For a fair comparison, we followed the same evaluation protocols used in [3]. The training data and testing data are separated based on the number of instances. 32 females and 33 males are randomly chosen for training, and for the rest of the instances, 28 females and 31 males are used for testing. In the training phase, we used 50% of the training data for training and 25% of the training data for validation.

In the testing phase, three different testing sets, namely TestSet-1, TestSet1-3, and TestSet1-4, are constructed from the testing data. TestSet-1 contains only a standing posture. TestSet1-3 combines the standing and walking postures. TestSet1-4 includes stands, walking, and interaction postures. Thus, TestSet1-3 and TestSet1-4 are more challenging than TestSet-1. This setting aims to assess the performance of the system in the three scenarios mentioned above: standing, walking, and interacting.

The statistics of the Gender dataset are summarized in Table 1. The numbers of samples in the training set and test set are listed in detail. The distribution of the male and female are almost equal in both the training set and testing sets.

B. EVALUATION METRIC

The first metric used for gender classification is the overall accuracy (ACC) which can be defined as (8).
TABLE 1: Statistic of Gender dataset [3]. The number of samples is counted based on each entity listed in the first column. The Instances row counts the number of the subjects’ videos. For abbreviation purposes, we used F for female and M for male in the table.

|          | Training Data | TestSet1 | TestSet1-3 | TestSet1-4 |
|----------|----------------|----------|------------|------------|
| Total    | 62,272         | 475      | 34,714     | 54,980     |
| # of Female | 32,320         | 225      | 19,160     | 26,387     |
| # of Male | 29,952         | 250      | 20,554     | 28,593     |
| # of Standing | 521            | 475      | 475        | 475        |
| # of Walking | 43,516         | 0        | 34,239     | 34,239     |
| # of Interaction | 14,985     | 0        | 0          | 20,266     |

Instances (F & M) | 32 & 33 | 28 & 31 | 28 & 31 | 28 & 31
Video Length | 353 | 8 | 290 | 317
Resolution | 256 x 256

\[
ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)
\]

where the \(TP\) represents the correct-classification of the male and \(TN\) represents the correct-classification of the female. The denominator \((TP + TN + FP + FN)\) covers all the predictions including the correct-classification and miss-classification.

Other typical evaluation metrics used in classification problem are also adopted in this work, such as Precision (P), Recall(R) and F1 score and they are defined as (9),(10), and (11) respectively.

\[
Precision = \frac{TP}{TP + FP} \quad (9)
\]

\[
Recall = \frac{TP}{TP + FN} \quad (10)
\]

\[
F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (11)
\]

C. IMPLEMENTATION DETAIL

We implemented two stages of model training, warm-up stage, and backbone training stage. During the warm-up stage, the two convolution branches are frozen and only the self-joint attention module and prediction head are trained. This stage is to let the self-joint attention module and prediction head have a good initialization. The warm-up stage goes through the whole training data using one epoch. After that, the backbone branches unfreeze the selected convolution layers, the C5 layer in the RGB branch, and C3-C5 in the depth branch, which are also mentioned in III-A. Both backbone branches use the pre-trained weights on ImageNet. The model is trained with stochastic gradient descent (SGD) with a batch size of 16 pairs of images, RGB, and depth. The training epoch is set to 10 and an early stop condition will take place when the training loss decreases less than 0.001 in the consecutive epoch. The learning rate is set to 0.01 for both the self-joint attention module and the prediction head in the warm-up stage and backbone training stage. The backbone learning rate is set to one-tenth of the learning rate since the backbone is doing fine-tuning. The learning rate decays exponentially from 0.01 in the last 5 epochs. Weights decay and momentum are set to 0.0001 and 0.9 in the SGD.

The whole training and testing are implemented in Python using Pytorch on a PC with Intel 1.8GHz CPU, NVIDIA GTX GPU, 8G RAM.

D. ABLATION STUDY

Backbone Comparison: We evaluated the performances of the model using five standard CNNs as the backbone. These CNNs include VGG19, VGG19 with batch normalization (VGG19bn), VGG16, VGG16 with batch normalization (VGG16bn), and ResNet50. The model is trained and tested using the protocol described in the Section IV-A and IV-C. The training curves of the model using different backbones are shown as in Figure 3. The training accuracy and training loss are measured at each epoch. The model is quickly converging after four epochs.

Table 2 shows the classification results using these standard CNNs as the backbone. The best and second-best performances are marked in bold and blue respectively for each evaluation metrics. The VGG19 and VGG16bn rank the best or second-best in all the testing sets. Note that the proposed model using different backbones has close performances on the three testing sets. It shows that the proposed model is effective to classify the person’s gender in different postures. Figure 4 shows the ROC curves of the model in the testing. It also shows that using the proposed model can classify the male class and female class with high precision.

Figure 5 reports the classification confidence distributions of the model using VGG16bn as the backbone. The red dot line represents the distribution of the false classifications while the blue dot line represents the distribution of the correct classifications. The distributions are calculated using the cumulative distribution function. Because TestSet1 has a limited number of data, the distribution curve is shown in Figure 5(a) is sparse. From the figures, We can see that the model has high confidence in the accurate prediction more than 85% while the model has lower confidence in the false prediction.

Impact of the fusion stage: In the proposed method, self-joint attention is applied to the features from the deep convolution layers. The top convolution layers have a larger receptive field than the bottom convolution layers. This is because the size of the feature maps reduces as we move towards the top layers. Hence the convolution filters will have larger receptive fields at the top layers. Therefore, the features from the top convolution layers can capture more semantic and global information though at the expense of spatial and local information. Hence, the self-joint attention module would have a different impact on the features from different stages of convolution layers. We modified VGG19 and VGG16bn backbone by inserting the self-joint attention module at different stages to compare the model’s perfor-
TABLE 2: Performance comparison of gender classification using different CNNs as the model’s backbone. The results are obtained using self-joint attention module. Each column is one individual testing. TestSet-1 only contains a standing pose. TestSet-1-3 combines standing and walking postures. TestSet1-4 includes all the testing samples. The four evaluation metrics are: Accuracy (ACC), Precision (P), Recall (R), and F1 score.

| BackBone | TestSet1 ACC | P | R | F1 | TestSet1-3 ACC | P | R | F1 | TestSet1-4 ACC | P | R | F1 |
|----------|--------------|---|---|----|----------------|---|---|----|----------------|---|---|----|
| VGG19    | 95.8         | 95.6 | 95.6 | 95.6 | 93.9          | 91.6 | 95.5 | 93.2 | 94           | 92.5 | 94.3 | 93.4 |
| VGG19bn  | 95           | 95.5 | 93.8 | 94.6 | 93.3          | 92.3 | 93.6 | 93  | 93.8         | 93.7 | 95  | 93.2 |
| VGG16    | 91.1         | 88.3 | 93.8 | 91  | 90.2          | 87.1 | 93.6 | 90.3 | 90.7         | 88.7 | 92.4 | 90.5 |
| VGG16bn  | 96.3         | 94.8 | 97.4 | 96.1 | 93.9          | 92.9 | 94.5 | 93.7 | 94.1         | 93.9 | 93.9 | 93.9 |
| ResNet50 | 92.5         | 92  | 92  | 92  | 92            | 89.9 | 93.8 | 91.8 | 92.4         | 91.8 | 94  | 92.9 |

These fusion variants of the model are trained using the same setting except for the selection of frozen blocks. This aims to fine-tune the convolution layers adapting to the task-specific domain before the fusion. When fusion is taken after Conv3, we only froze the Conv1, and Conv4 and Conv5 are trained from scratch. Similarly, when the fusion is taken after Conv4, we froze Conv1 and Conv2, and Conv5 is trained from scratch. The rest of the convolution layers will be fine-tuned. These variants are also tested using the same protocol.

From Table 3, we can see that early-stage fusions on Conv4 and Conv3 degrade the model’s performance compared to late-stage fusion on Conv5 on both backbones. In the case of using VGG19 as the backbone, when fusing on Conv4, the accuracy reduces around 4% or 6% on all the test sets. But when fusing on Conv3, the accuracy is around 51% and the precision is 0 in all the testing sets which means the models have no discrimination capacity to classify the male class and only can recognize the female class. Thus the Precision, Recall, and F1 score are 0. In the case of using VGG16bn as the backbone, fusion on Conv3 and Conv4 have the same accuracy and precision and they are very low.

These results show that the self-attention module on the
late-stage can focus more on the important part of the features while on the early-stage the self-attention module adds attention to the non-important place on the features. This is explicit when we visualize the heatmaps of the predictions in the testing. Figure 6(a-c) show the heatmaps on the same testing samples based using the VGG19 backbone and the corresponding different fusion stages. In the late-stage fusion like Conv5, the heatmaps have high responses on the discriminative part of the person. In contrast, in the early-stage fusion like Conv3 and Conv4, the heatmaps usually have high responses on the less discriminative part or even the background and thus models’ performance degrades. In terms of computation overhead, the results in the last column of Table 3 show the computation cost is almost the same using different fusion stages. This is because the proposed self-joint attention module is lightweight.

Impact of Self-joint attention: We also investigated the effectiveness of using self-joint attention by comparing the model with and without the self-attention module. We continued using the VGG19 and VGG16bn networks as backbones for comparison. In the case of without self-joint attention module, we fused the features from Conv5 using element-wise addition with a softmax normalization to add the features. The training and testing are using the same protocols mentioned in Section IV-C.

In Table 3, the performances of the model without self-joint attention module are listed in rows marked as "Conv5 W A". From the comparison between "Conv5 W A" and "Conv5 WA", the self-joint attention module improves the model’s performances on the three testing sets, especially when using VGG16bn as the backbone. Note that when the model uses VGG16bn as the backbone and without the self-joint attention module, it has the same performance as when the model is doing early-stage fusion. The model is unable to recognize the two different classes. Inversely, this is proved that the proposed self-joint attention module does enhance the model’s discrimination capability for person gender classification.

We can find more impact of the self-joint attention module from the heatmaps in Figure 6. The model using the self-joint attention module as shown in Figure 6(a) adds more concentration on the body part than the model without the self-joint attention module as shown in Figure 6(d).

Impact of selective fusion: We trained and tested the model using the self-joint attention module without the selective fusion scheme described in III-C to look into the impact of the selective fusion. The training and testing follow the same
protocols. Their testing results are depicted in the "Conv5 WS" rows in the Table 3.

Compared to the results between "Conv5 WS" and "Conv5", the model using VGG19 as the backbone has close performances on the three testing sets. While using VGG16bn as the backbone, the selective fusion scheme is better than the without selective fusion scheme in terms of accuracy, precision, and F1 score, especially in TestSet1. The comparison shows adding the learnable selective fusion in the self-attention module improves the model's performance.

E. STATE-OF-THE-ART COMPARISON

We compared our method with state-of-the-art methods on the same dataset following the same experimental protocols described in IV-A. We chose the two best performances of the model using VGG19 and VGG16bn as backbones as the results described in Section IV-D.

The performance comparison is depicted in Table 4. We can observe that our approach reaches the best accuracy of 96.3%, 93.9%, and 94.1% for the three testing sets respectively. Compared to the state-of-the-art methods, our method improves the accuracy with an increment of around 5.2% on the simplest TestSet-1 and a large increment around 7.5% and 8.7% on the two other challenging test sets (TestSet1-3 and TestSet1-4), respectively. Our method significantly boosts the performance on the challenging testing sets including the walking and interaction postures. Also, compared to the others, our method has a close accuracy on the three testing sets while the other methods have a significant gap between the testing sets. This shows our method is robust and invariant to the change of posture.

F. STATISTICAL SIGNIFICANCE TEST

A statistical significance test is conducted to examine the performance of the proposed method and the state-of-the-art methods. The statistic significance test is designed to quantify the likelihood of the performance scores, which refers to accuracy in this work, being observed given the assumption that they are were drawn from the same distribution. If this assumption, also called the null hypothesis, is rejected, it implies that the difference in the scores is statistically significant. The most commonly used statistical significance test is the paired Student’s t-test as described in (12). The null hypothesis in this test is that there is no difference between the performance of the two models.

$$t_{1,2} = \frac{\bar{S}_1 - \bar{S}_2}{\bar{\sigma}_{pool}}$$

$$\bar{\sigma}_{pool} = \sqrt{\frac{(\sigma_1)^2 + (\sigma_2)^2}{2}}$$

where $\bar{S}_1$ and $\bar{S}_2$ are the means of the two models’ performance. $\sigma_1$ and $\sigma_2$ are their standard deviation.

Table 5 shows the t values of the comparison between the proposed method and the state-of-the-art method. The values are calculated at the 0.05 significance level. Considering the critical value $t_{critical} = 4.303$, we can see that all the t values are larger than 4.303. Also, all the p-values are smaller than the significance level. The null hypothesis is rejected. Therefore, the results show that the proposed model
TABLE 4: Performance comparison of gender classification with other methods. Our proposed self-joint attention (SJA) method achieves the best performance in all the testing TestSets.

| TestSet1 | TestSet1-3 | TestSet1-4 |
|----------|------------|------------|
| ACC      | P | R | F1 | ACC | P | R | F1 | ACC | P | R | F1 |
| HOG32 [15] | 84.78 | \ | \ | \ | 70.55 | \ | \ | \ | 70.70 | \ | \ | \ |
| HOG64 [15] | 86.27 | \ | 86.10 | \ | 74.10 | \ | 86.10 | \ | 74.36 | \ | 86.30 | \ |
| CRNN [16] | 86.64 | \ | 86.07 | \ | 83.46 | \ | 85.28 | \ | 83.73 | \ | 86.19 | \ |
| HMP [17] | 88.07 | \ | 87.91 | \ | 86.41 | \ | 85.29 | \ | 85.29 | \ | 86.39 | \ |
| Tessellation [3] | 91.07 | \ | 90.87 | \ | 93.61 | \ | 93.42 | \ | 93.54 | \ | 93.72 | \ |
| SJA (VGG19) | 95.8 95.6 95.6 95.6 | 93.9 91.6 95.5 93.2 | 93.5 92.5 94.3 93.4 | 94.1 93.9 93.9 93.9 |

TABLE 5: Statistic significance test compares the performances of the proposed self-joint attention model (SJA) and the state-of-the-art models.

| t(SJA, Tessellation [3]) | \ | \ | \ | \ | \ | \ | \ | \ | \ | \ | \ |
| t(SJA, HMP [17]) | 45.30 | 0.0005 | 4.303 | 45.25 | 0.0006 | 4.303 | 45.30 | 0.0005 | 4.303 |
| t(SJA, CRNN [16]) | 34.63 | 0.0008 | 4.303 | 34.62 | 0.0008 | 4.303 | 34.63 | 0.0008 | 4.303 |
| t(SJA, HOG64 [15]) | 5 | 0.0368 | 4.303 | 5 | 0.0369 | 4.303 | 5 | 0.0368 | 4.303 |

TABLE 6: Evaluate the model based on person orientation. In each subset, the persons are in the same orientation.

| Subset | Samples | Orientation (degree) | ACC |
|--------|---------|----------------------|-----|
| Ortt1 | 28 | 31 | 0 | 96.88 |
| Ortt2 | 28 | 31 | 45 | 96.88 |
| Ortt3 | 28 | 31 | 90 | 92.19 |
| Ortt4 | 28 | 31 | 135 | 98.44 |
| Ortt5 | 28 | 31 | 180 | 96.88 |
| Ortt6 | 28 | 31 | 225 | 96.88 |
| Ortt7 | 28 | 31 | 270 | 96.88 |
| Ortt8 | 28 | 31 | 315 | 96.88 |

and the state-of-the-art are performed differently and their performances are statistically significant.

G. PERSON ORIENTATION TEST

We sampled eight subsets, Ortt1 to Ortt8, from the testing sets to evaluate the model’s sensitivity to the orientation of the person. The eight subsets contain the person in eight different orientations, as described in Table 6, respectively. Each subset contains one sample from 28 females and 31 males standing in the same orientation. We evaluated the model with VGG16bn as the backbone using these subsets. The results show that the accuracies of most of the subsets are the same around 96.88%. These observations evidence that our proposed model is robust to the change of the person’s orientation.

V. CONCLUSION

In this paper, we proposed a self-joint attention method for gender classification of persons captured by RGB-D camera and exhibiting challenging scenarios of different views, poses, and scales. The method uses an attention mechanism to fuse the features from two different modalities through encoding the inter-dependent information between them so that the model can learn more discriminative features. We conducted a thorough ablation study to set the parameters of our system to evaluate the performance. The proposed method outperforms the state-of-art in all three different testing sets with significant margins of 4.7%, 7.5%, and 8.7% respectively on a public RGB-D gender dataset.

REFERENCES

[1] Gil Levi and Tal Hassner, “Age and gender classification using convolutional neural networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2015, pp. 34–42.
[2] M Annalakshmi, S Mohamed Mansoor Roomi, and A Sheikh Naveedh, “A hybrid technique for gender classification with slbp and hog features,” Cluster Computing, vol. 22, no. 1, pp. 11–20, 2019.
[3] T. Linder, S. Wehner, and K.O. Arras, “Real-time full-body human gender recognition in (rgb)-d data,” in IEEE ICRA, 2015.
[4] Feng Lin, Yingxiao Wu, Yan Zhuang, Xi Long, and Wenyao Xu, “Human gender classification: a review,” International Journal of Biometrics, vol. 8, no. 3-4, pp. 275–300, 2016.
[5] Beom Kwon and Sanghoon Lee, “Joint swing energy for skeleton-based gender classification,” IEEE Access, vol. 9, pp. 28334–28348, 2021.
[6] Balakrishna Gudla, Srinivasas Rao Chalamala, and Santosh Kumar Jami, “Local binary patterns for gender classification,” in 2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS). IEEE, 2015, pp. 19–22.
[7] Joseph Lemley, Sami Abdul-Wahid, Dipayan Banakar, and Razvan Andonie, “Comparison of recent machine learning techniques for gender recognition from facial images,” MAICS, vol. 10, pp. 97–102, 2016.
[8] Sandeep Kumar, Sukhwinder Singh, and Jagdish Kumar, “Gender classification using machine learning with multi-feature method,” in 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2019, pp. 0648–0653.
[9] E-SM El-Alfy and Amer G Binsaooda, “Automated gait-based gender identification using fuzzy local binary patterns with tuned parameters,” Journal of Ambient Intelligence and Humanized Computing, vol. 10, no. 7, pp. 2495–2504, 2019.
[10] E. Boutellaa, A. Hadid, M. Bengherabi, and Samy A.A, “On the use of kinect depth data for identity, gender and ethnicity classification from facial images,” PRL, vol. 68, pp. 270–277, 2015.
[11] Yuechen Yu, Yilei Xiong, Weilin Huang, and Matthew R Scott, “Deformable siamese attention detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6728–6737.
[12] Nian Liu, Ni Zhang, and Junwei Han, “Learning selective self-mutual attention for rgb-d saliency detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13756–13765.
[13] Lahoucine Ballihi, Boulbaba Ben Amor, Mohaned Daoudi, Anuj Srivastava, and Driss Aboutajdine, “Geometric based 3d facial gender classification,” in 2012 5th International Symposium on Communications, Control and Signal Processing. IEEE, 2012, pp. 1–5.
[14] S. Azzakhnni, Lahoucine B., and Driss A., “Machine perception in gender recognition using rgb-d sensors,” in IEEE AICCSA, 2016.
[15] L. Spinello and K.O. Arras, “People detection in rgb-d data,” in IEEE IROS, 2011.
[16] R. Socher, B. Huval, B. Bath, Christopher D Manning, and A.Y. Ng, “Convolutional-recursive deep learning for 3d object classification,” in NIPS, 2012.
[17] L. Bo, X. Ren, and D. Fox, “Learning hierarchical sparse features for rgb-d object recognition,” IJRR, vol. 33, no. 4, pp. 581–599, 2014.
[18] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” in International Conference on Learning Representations, 2015.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in Advances in neural information processing systems, 2017, pp. 5998–6008.

***