Pattern Analysis and Applications (2024) 27:49
https://doi.org/10.1007/s10044-024-01270-3

ORIGINAL ARTICLE

Proxemics-net++: classification of human interactions in still images

Isabel Jiménez-Velasco[1] · Jorge Zafra-Palma[1,2] · Rafael Muñoz-Salinas[1,2] · Manuel J. Marín-Jiménez[1,2]

Received: 27 October 2023 / Accepted: 11 April 2024

Abstract
Human interaction recognition (HIR) is a significant challenge in computer vision that focuses on identifying human interactions in images and videos. HIR presents a great complexity due to factors such as pose diversity, varying scene conditions, or the presence of multiple individuals. Recent research has explored different approaches to address it, with an increasing emphasis on human pose estimation. In this work, we propose Proxemics-Net++, an extension of the Proxemics-Net model, capable of addressing the problem of recognizing human interactions in images through two different tasks: the identification of the types of "touch codes" or proxemics and the identification of the type of social relationship between pairs. To achieve this, we use RGB and body pose information together with the state-of-the-art deep learning architecture, ConvNeXt, as the backbone. We performed an ablative analysis to understand how the combination of RGB and body pose information affects these two tasks. Experimental results show that body pose information contributes significantly to proxemic recognition (first task) as it allows to improve the existing state of the art, while its contribution in the classification of social relations (second task) is limited due to the ambiguity of labelling in this problem, resulting in RGB information being more influential in this task.

Keywords Human interactions · Proxemics · Social relations · Human pose estimation · Deep learning

1 Introduction

Human activity recognition (HAR) is one of the most important and challenging problems in computer vision, which aims to recognize activities present in images or videos automatically. In particular, Human Interaction Recognition (HIR) constitutes a subset of HAR and focuses on distinguishing human-to-human interactions within visual data [1], such as handshakes, hugs, conversations or even what types of physical contact or Proxemics [2] exist between pairs of people (hand-hand, hand-shoulder, etc.). The latter provides very relevant information to determine the type of social interaction and the interpersonal relationships between the members present in the interaction since the type of physical contact will vary greatly depending on whether they are acquaintances, friends or co-workers.

Recognizing human-human interactions in images and videos is a fundamental challenge in computer vision and deep learning. Its importance extends to numerous real-world applications, including human-computer interaction, surveillance systems, autonomous vehicles, and other fields [3–6].

HIR is challenging due to the complex postures of human beings, the number of people in the scene, and specific challenges, such as illumination variations, clutter, occlusions, and background diversity. For example, if we look at Fig. 1, which shows people interacting differently, how would the reader classify these interactions? In the first two, we can see two couples interacting physically, but what kind of physical contact can we clearly observe: hand-hand, hand-shoulder, etc.? In the following two images, can the reader determine with complete certainty from visual information alone whether the pairs of people are friends, family members, or

✉ Isabel Jiménez-Velasco
isajimenez@uco.es

Jorge Zafra-Palma
jzafra@uco.es

Rafael Muñoz-Salinas
rmsalinas@uco.es

Manuel J. Marín-Jiménez
mjmarin@uco.es

[1] Department of Computing and Numerical Analysis, University of Córdoba, Córdoba, Spain

[2] Maimonides Institute for Biomedical Research of Córdoba (IMIBIC), Córdoba, Spain

Published online: 29 April 2024

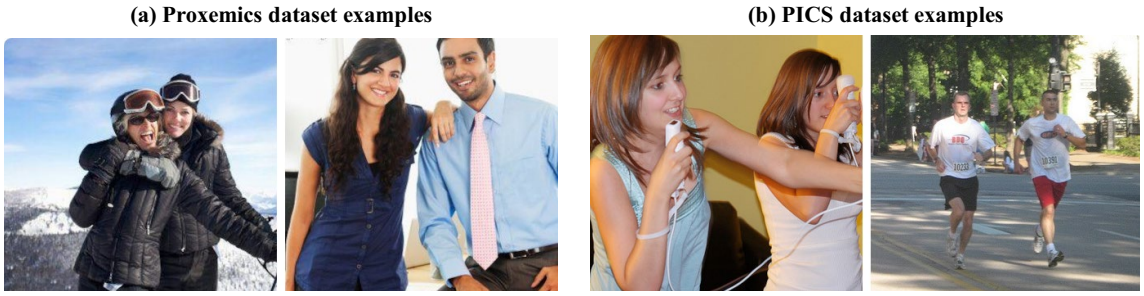

Fig. 1 Examples of human-human interactions. Could the reader indicate what kind of contact exists between the pairs in the examples in (**a**)? What kind of relationship exists between the people in (**b**)? These images illustrate the great complexity inherent in the problem of recognizing human interactions in images. The images in (**a**) highlight situations where it is confusing to determine the type of physical contact (hand-elbow, hand-shoulder, elbow-shoulder, etc.) due to clothing and partial occlusion. In (**b**), the images show ambiguity in determining the type of social relationship between individuals (family, friends, co-workers, etc.) in the absence of additional context

co-workers? Which of those tasks is easier for the reader? As we can see, these images reflect the great complexity of the problem of recognizing human interactions in images and videos.

In recent years, the HIR problem has been addressed from different perspectives. For example, Yang et al. [2] address this problem by making use of Proxemics, a branch of anthropology that studies interactions between people by analyzing their proximity and how they physically touch each other (hand-hand, hand-shoulder, etc.). On the other hand, in [7], the authors address the problem by focusing on the visual recognition of social relationships in images (family, friends, professionals, etc.).

Human interactions are characterized by several key elements, such as context, scene, and pose [8]. Of these elements, it can be argued that human poses are central to identifying an interaction, as context and scene can change considerably, while pose must remain constant as members of a recognizable set of interaction categories.

Human pose estimation has become an increasingly popular research topic in the last decade due to its application in various areas, including human interaction recognition in images. This has led to recent work demonstrating the important contribution of human pose to this problem [8–10].

In addition, the growing popularity of Computer Vision and the rapid advancement of powerful Deep Learning techniques have given rise to much newer methods and approaches such as Vision Transformers [11] or ConvNeXt [12].

The main objective of this work is to address the problem of detecting human interactions in images from two different tasks. From a lower level task in which we focus on the detection and classification of the type of physical interactions between individuals (proxemics) to a higher level task in which we focus on the classification of the type of social interaction between pairs of people in which there may or may not be physical interaction.

In a preliminary version of this work [13], we proposed a new model, coined Proxemics-Net, and investigated the performance of two state-of-the-art deep learning architectures, ConvNeXt [12] and Visual Transformers [11] (as backbones) on the problem of proxemic recognition using only RGB information (not pose information). We showed experimental results that outperformed the existing state-of-the-art and demonstrated that the two state-of-the-art deep learning models help in the proxemics recognition problem using only RGB information, with the ConvNeXt architecture being the best-performing backbone.

The main new contributions of this work are as follows:

- We propose Proxemics-Net++, an extension of the Proxemics-Net model that combines the ConvNeXt architecture with RGB and body pose information to address our two proposed tasks for recognizing human interactions in images: categorizing physical interactions (proxemics) and social interactions between pairs.
- We propose a body pose representation from the information obtained by the 3D pose estimator DensePose [14] after being applied to our datasets.
- We perform an ablative study to analyze how the type of information employed (RGB and Pose) and their combination influence the human interaction recognition problem.
- We show experimental results that outperform the existing state of the art on the Proxemics dataset. This indicates that body pose information and state-of-the-art deep learning architectures contribute significantly to our first task, which focuses on proxemics recognition in images.
- Our experiments reveal that, unlike RGB data, body pose information is insufficient for our second task,

which aims to identify social relationships between pairs where there may or may not be physical contact. As seen in the results (Sect. 5.2), the inherent complexity and ambiguity of the problem indicate the need for a different architecture and additional information.

The remainder of this paper is structured as follows. Section 2 presents the related work, and then, in Sect. 3, we describe the proposed new Proxemics-Net++ model. Later, we will explain all the experiments' characteristics and the implementation details (Sect. 4). Then, we will show and comment on the results of all the experiments performed (Sect. 5), and finally, we will finish with some conclusions and future work (Sect. 6).

2 Related work

Over the last few years, human interaction recognition in images has received increasing attention from the research community.

This is a difficult problem due to several reasons. First, some interactions, such as kissing or shaking hands, involve only two people, while others, such as dining or partying, may involve a larger number of individuals. Second, there is no restriction on the characteristics of the images, such as camera position, lighting, or clutter, which can vary considerably and lead to ambiguity or occlusions of the people involved. Finally, the presence of unrelated people in the scene complicates the identification of the actual interaction.

To address the problem of human-to-human interaction detection, the anthropologist Hall [15] introduced the concept of *proxemics*, a categorization of human individual interactions based on spatial distances. In 2012, Yang et al. [2] characterized Proxemics as the problem of recognizing how people physically touch each other and called "touch codes" to each type of interaction or proxemics. Yang et al. identified six dominant "touch codes": Hand-Hand, Hand-Shoulder, Shoulder-Shoulder, Hand-Torso, Hand-Elbow, and Elbow-Shoulder. The authors of this study claimed that using specific detection models was the best way to address the problem of proxemics recognition because other alternatives, such as pose estimation, were significantly affected by ambiguity and occlusion when there was physical interaction between people.

To address the challenges arising from ambiguity and occlusion during physical interactions between individuals, Xiao et al. [16] expanded the concept of the "touch code" by incorporating additional information. Specifically, they introduced a complete set of "immediacy" cues, encompassing physical contact, relative distances, body leaning direction, eye contact, and standing orientation. The authors developed a Deep Multi-task Recurrent Neural Network (RNN). This neural network was designed to model the intricate correlations between these immediacy cues and human pose estimation. This novel model demonstrated significant advancements over the state-of-the-art results achieved by [2].

In 2017, motivated by the limitations of existing people detection methods related to speed, efficiency, and performance, particularly in scenarios involving unknown scales and orientations, occlusion, and ambiguity in identifying body parts, Jiang et al. [17] proposed a new approach. This new method aimed to segment individual humans and label their body parts, including arms, legs, torso, and head, by assembling regions. Consequently, this new approach improved the state of the art on proxemics recognition.

The problem of human interaction recognition can also be addressed from other approaches. For example, Li et al. [7] focus on visual recognition of social relationships in images. They proposed a Dual-glance model for social relationship recognition, where the first glance fixates on the person of interest, and the second glance deploys an attention mechanism to exploit contextual cues.

Several notable approaches have been introduced in recent years to address the problem of recognizing social relations from images, mainly focusing on graph-based models. In 2019, the Multi-Granularity Reasoning (MGR) framework was proposed by Zhang et al. [18]. MGR emphasizes the importance of capturing social relationships through a combination of different sources of information, including the Person-Object Graph (POG) to model actions between people and objects and the Person-Position Graph (PPG) to represent interactions between matched individuals. Thus, this framework combines global knowledge, regional features, and interactions between people and objects to improve the recognition of social relationships.

Goel et al. [19] presented a Social Relationship Graph Generation Network (SRG-GN). SRG-GN stands out as an end-to-end trainable neural network capable of generating a Social Relationship Graph from input images. This innovation, which uses memory cells as Gated Recurrent Units (GRUs), offers a dynamic approach to iteratively update social relationship states in a graph. In this case, the network integrates scene context and attributes, which provides additional information on social relationships and their attributes. Li et al. [20] proposed a graph relational reasoning network (GRN) for social relation recognition in images. Unlike other methods developed to date, it considered all social relations in an image together by constructing a graph of social relations, rather than independently. This approach not only improved accuracy and efficiency, but also managed the logical constraints

between different types of social relations by constructing several virtual relation graphs.

In 2021, to address the limitations of previous methods based on multiple relationship graphs, Li et al. [21] introduced a novel method known as Hybrid-Features Social Relation Graph Reasoning (HF-SRGR). Their approach focuses on capturing dependencies among multiple relationships while incorporating contextual information. In particular, HF-SRGR constructs a graph in which each node represents a relationship, with the addition of a scene node.

Yang et al. [22] proposed a Gaze-Aware Graph Convolutional Network (GA-GCN) for social relation recognition, which aims to discover context-aware social relation inference with gaze-aware attention. Finally, Sousa et al. [23] proposed a novel approach based on a Social Graph Network (SGN) capable of interpreting relationships from three different domains (individual, relative, and general information). Additionally, prior knowledge was also considered since how humans differentiate relationships is deeply associated with appearance attributes such as age, gender, clothing, emotion, and pose. This new approach improved the state of the art on social relation recognition.

In this work, we focus on the two tasks mentioned above related to human-to-human interaction in still images: proxemics recognition based on the classification of touch codes and high-level social interaction classification. Instead of using elaborated graph-based models, we are interested in studying how feed-forward neural architectures based on Convolutional Neural Networks can be applied to those tasks.

On the other hand, in the last decade, human pose estimation has also become an increasingly popular research topic due to its application in various areas, including human interaction recognition in images. That is why, recently, several researchers have addressed the problem of human interaction recognition using body pose estimation, demonstrating a significant contribution to this problem.

In [24], the authors presented a method for recognizing human interactions from videos by combining high-level features computed by a Convolutional Neural Network pre-trained on Imagenet, with articulated body joints as low-level features. Gokhan et al.'s work [8] proposed a Multi-stream Convolutional Neural Network architecture, mainly focused on pose information. Specifically, several pose-based representations were formulated, showing that paying more attention to poses positively affects human interaction recognition. In [9], a novel framework for human interaction recognition, which considers both the implicit and explicit representations of human behavior, is presented. In [25], the authors propose a deep learning approach that recognizes humans and their social interactions in a 3D space from visual cues.

In the early version of this work [13], we presented Proxemics-Net, a model that uses RGB image information and advanced deep learning architectures, particularly ConvNeXt and Visual Transformers, for proxemic classification. Experiments on the existing Proxemics dataset showed that these architectures significantly improved performance on this task, with ConvNeXt being the most effective.

Therefore, in this new work, we propose Proxemics-Net++, an extension of the Proxemics-Net model, able to address the problem of human interaction recognition in images from two different tasks: by identifying the type of proxemics as well as the identification of the type of social relationship between couples. For this purpose, we combine RGB and body pose information with the ConvNeXt architecture (best backbone obtained). In this way, we also analyze how the human pose information contributes to each task and whether it reduces the ambiguity present in the images.

3 Proposed method

In this section, we will first introduce the original model proposed in our previous work (Sect. 3.1) and briefly discuss the ConvNeXt architecture that we have used as a backbone (Sect. 3.2). Then, we will summarize DensePose, the pose estimator we have selected, and the body pose representation we generate from its output (Sect. 3.3). Finally, we will explain the new Proxemics-Net++ model proposed in this work (Sect. 3.4).

3.1 Overview of the base model: Proxemics-Net

In our previous work [13], a new model called Proxemics-Net was proposed (see blue branches in Fig. 2). This model was based on two different deep networks that had been previously pre-trained: ConvNeXt [12] and Vision Transformers (ViT) [11].

The Proxemics-Net model has three inputs. Two inputs corresponding to the RGB clipping of each of the individuals composing a pair (*p0_branch*) and (*p1_branch*) and a third input corresponding to the RGB clipping showing the pair to be classified (*pair_branch*). The three input branches receive RGB images of 224×224 resolution.

All three branches of this base model used a common backbone to process the inputs (ConvNeXt [12] or Vision Transformers (ViT) [11]). The results of the three branches are merged through a concatenation layer and then passed through a fully connected layer, which predicts the proxemic classification of the input samples.

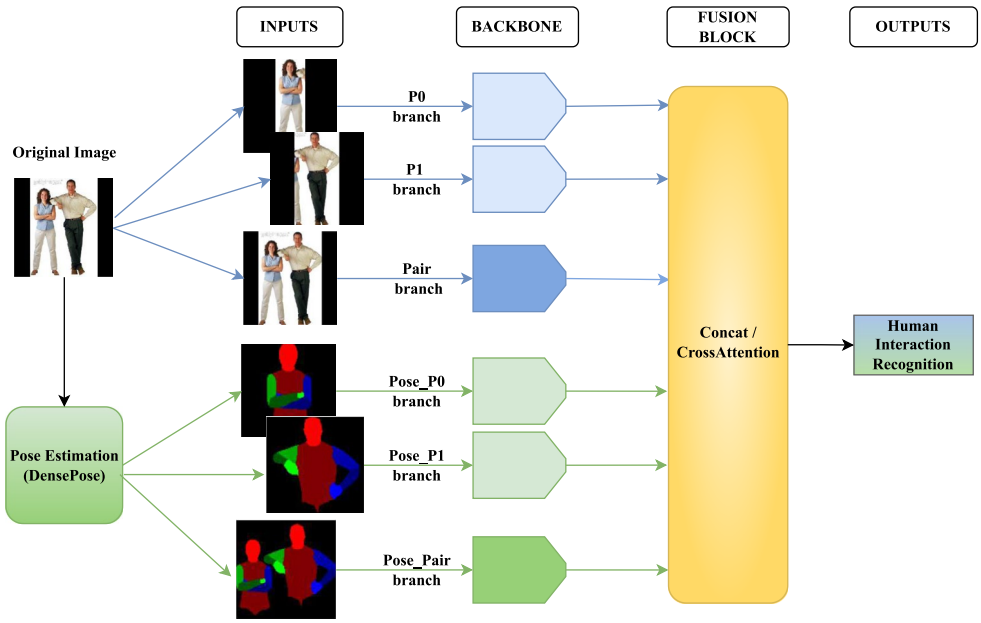

Fig. 2 Our Proxemics-Net++ model. It consists of six inputs: three branches for the RGB information of the couple and the individuals that compose it (blue branches) and another three branches for the body pose representation of the two individuals and the couple (green branches). All branches have the same type of backbone (Base or Large). The outputs of these six branches are passed to a Fusion Block, which can be of two types: Concatenation fusion or CrossAttention fusion (see Fig. 4). Finally, the type of human interaction (proxemics or social relationship) of the input samples is predicted (best viewed in colour) (color figure online)

3.2 ConvNeXt-based backbone

The ConvNeXt model was proposed by Liu et al. [12] and is a pure convolutional model (ConvNet) constructed entirely from standard ConvNet modules and inspired by the design of Vision Transformers [11].

ConvNeXt models are built on a foundation of conventional neural network components, including depthwise convolutions and layer normalization. Unlike recent architectures such as Vision Transformers (ViT), ConvNeXt models do not incorporate self-attention mechanisms or hybrid approaches. Instead, in their development, the authors adopt a stepwise approach, starting with a basic architecture and progressively enhancing it. This process aligns with certain design principles observed in ViT, such as the emphasis on layer normalization and a deep and scalable network structure (network capable of managing different types and sizes of data), which are crucial for performance and efficiency. In the process, they developed a family of models named ConvNeXt, achieving high performance on the ImageNet-1k and ImageNet-21k datasets [26].

However, it should be noted that the ConvNeXt model has not been developed in isolation, but rather as an evolution of other transformers, in particular the Sliding Window Transformer (Swin) [27]. The Swin Transformer distinguishes itself by processing images at a higher level of granularity thanks to its self-attention mechanism, which effectively solves the scalability problems present in the vanilla Vision Transformers.

In our previous work [13], a comparison was made between the ConvNeXt and ViT architectures, resulting in ConvNeXt as the best backbone for our proxemics recognition problem. However, since ConvNeXt is an adaptation of Swin, we have extended our comparison and included the Swin Transformer as a possible backbone, in order to provide a fairer and more complete comparison between these state-of-the-art architectures and to be able to conclude with certainty which architecture is better for our problem.

Table 1 shows the comparison between the best results obtained with the selected Swin Transformer variants and the results obtained in our previous work [13]. Specifically, we have conducted a series of experiments using the Swin Transformer in two variants (Tiny [1] and Base [2]) with the Proxemics dataset and RGB information, similar to what we did in [13].

[1] microsoft/swin-tiny-patch4-window7-224: https://huggingface.co/microsoft/swin-tiny-patch4-window7-224.

[2] microsoft/swin-base-patch4-window7-224-in22k: https://huggingface.co/microsoft/swin-base-patch4-window7-224-in22k.

Table 1 Results of the best models obtained for each of the three proposed backbone types on the Proxemics dataset with RGB information

Backbone	HH	HS	SS	HT	HE	ES	mAP (Set1-Set2)
ViT	45.5	**55.7**	46.4	76.6	56	51.9	55.7
SwinTransformer_Tiny	51.9	50.5	46.1	73.8	52.3	48.3	53.8
SwinTransformer_Base	**59.3**	54.1	45.0	83.5	54.4	52.7	58.2
ConvNeXt_Base	54.1	51.7	59.3	82.3	57.2	54.2	59.8
ConvNeXt_Large	57.7	54.9	**59.3**	**85.3**	**60.3**	**64.9**	**63.7**

The results (average of set1 and set2) show that the model incorporating the ConvNeXt network as a backbone obtains the best results. The %AP results for each type of proxemics and the %mAP are shown.

Bold values represent the best result obtained in each of the six labels and in mAP (among allthe models compared)

Our results indicate that while the Swin Transformer models perform better than ViT, they do not outperform the results of ConvNeXt, thus confirming ConvNeXt as the best backbone for our problem.

3.3 Body pose representation: DensePose

DensePose [14] is a human pose estimator that aims to map all human pixels in a 2D RGB image to the 3D surface of the human body.

Within the DensePose project, two main approaches are used: chart-based dense pose estimation and continuous surface embeddings.

The goal of chart-based DensePose methods is to establish dense correspondences between image pixels and the 3D object mesh by splitting the latter into charts and estimating, for each pixel, the corresponding chart index (I) and local chart coordinates (U, V). Specifically, the human body is divided into 24 parts.

For each detected human, the model predicts its coarse segmentation S (with 2 or 15 channels representing foreground/background or background plus 14 predefined body parts), fine segmentation I (with 25 channels representing background plus 24 predefined body parts), and local chart coordinates U and V.

For this work, we have used the DensePose estimator to obtain the pose of all the persons present in the images of our datasets. However, we have selected only the returned I-map of each detected person and made some changes. First, of the 24 predefined body parts, we have discarded the parts corresponding to the legs and kept only the parts corresponding to the head, torso, and arms since the characteristic details of physical interactions and social relationships are most often visually located in the upper body, with the lower body being unnecessary. Once the lower part of the human body has been discarded, we have grouped the different body parts in a 224×224 TLR map (Torso, Left_arm, Right_arm). Specifically, the T channel corresponds to the Red channel of the map and includes pixels of the head and torso. Meanwhile, the L and R channels correspond to Blue and Green, respectively. The L channel contains pixels of the left arm (arm, forearm, and hand), and the R channel contains pixels of the right arm (arm, forearm, and hand). SpecificallWithin each channel, we differentiate each grouped subpart with a different pixel value (see green box in Fig. 3). In this way, we obtain a body pose representation of each detected person's parts and the pair to be evaluated.

3.4 Our proposed model: Proxemics-Net++

In this work, we propose Proxemics-Net++, a model that extends Proxemics-Net with some variations. In particular, we have added three new inputs for the body pose representation we obtained from the pose estimator DensePose.

Thus, this new model has six inputs, the three inputs it already had for RGB and three new inputs for pose (see Fig. 2).

These three new inputs maintain the same structure and format as the three RGB inputs. That is, two inputs corresponding to the body pose of each of the individuals composing a pair (*pose_p0 branch*) and (*pose_p1 branch*) and a third input corresponding to the clipping showing the body pose of the pair to be classified (*pose_pair branch*). These three input branches receive images of 224×224 resolution.

The six branches of the model have the same type of backbone. In this case, we have only made use of the ConvNeXt pre-trained deep network as a backbone since it was the one that obtained the best results in the previous work [13]. The backbone of each branch is responsible for extracting the characteristics of the corresponding input.

The results of the six branches are combined in a Fusion Block consisting of a Concatenation Fusion or a CrossAttention Fusion (see Fig. 4).

In the Concatenation Fusion Block (see Fig. 4–left image), the input branches are combined through a concatenation layer and passed through a fully connected layer that predicts the type of human interaction (proxemics or social relationship, depending on the particular task) of the input samples.

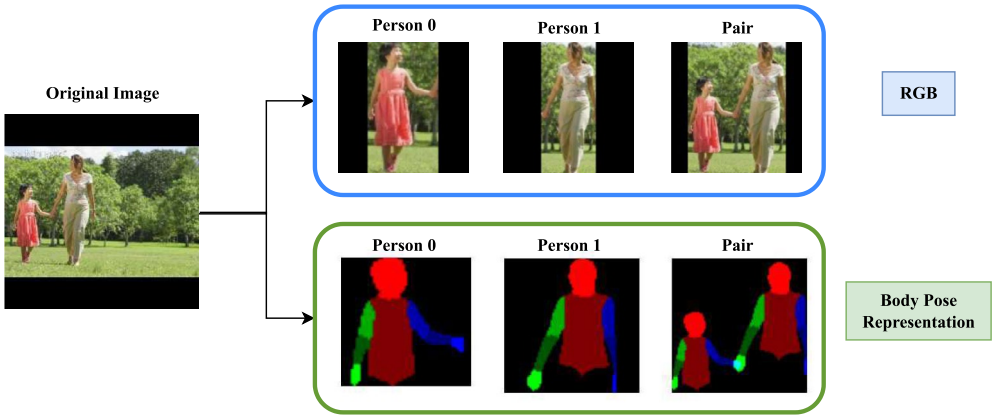

Fig. 3 Input data for Proxemics-Net++. Given the target image, the first two image crops ('Person 0' and 'Person 1') correspond to the individual clippings of the members of the pair, and the third image crop ('Pair'). Specifically, the blue rectangle shows the corresponding RGB clippings and the green rectangle shows the corresponding clippings of the body pose representation. All clipping images have a resolution of 224 × 224 pixels (color figure online)

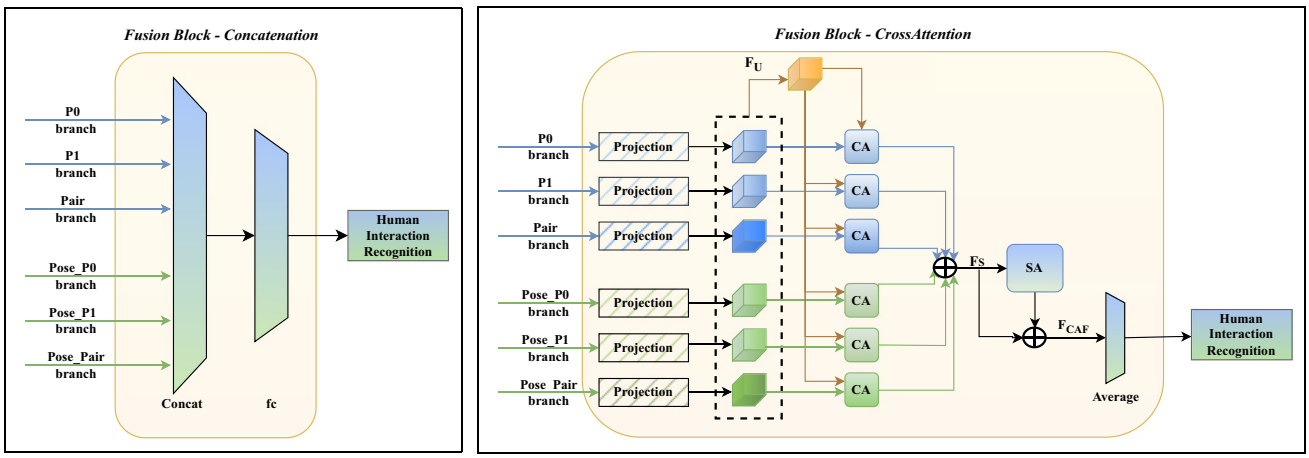

Fig. 4 Fusion blocks types implemented in Proxemics-Net++. In the Concatenation Fusion Block (left image), the results of the six branches are combined through a concatenation layer and a fully connected layer that predicts the human interaction type of the input samples. In the CrossAttention Fusion Block (right image), the outputs of the six branches pass through a Projection module that takes the features of each branch and projects them into a common 512 D feature space. Subsequently, a CrossAttention Fusion, consisting of CrossAttention (CA) and SelfAttention (SA) for feature fusion, is applied to all these outputs, following the same methodology implemented in [28]

In the CrossAttention Fusion Block (see Fig. 4–right image), we have implemented the same CrossAttention fusion method described in [28]. Cross-attention allows to analyse and understand correlations between different inputs (RGB and Pose images, for example). This approach does not simply process each input in isolation, but also learns to identify patterns and correlations between them by "paying attention" to how they relate to each other. In tasks that demand a thorough comprehension of the data, such as Human Interactions Recognition, the model can capture complex relationships.

Thus, following the fusion method described in [28], each of the six branches passes through a Projection module composed of a linear layer followed by batch normalisation, a *tanH* activation function and another linear layer. In this way, it takes the input features of each branch and maps them into a common 512 D feature space.

Once the features are projected, they are all stacked as $F_U \in \mathbb{R}^{6\times512}$, and six cross-attention modules (CA) are generated, one for each branch. These modules consist of a MultiHeadAttention layer. In this case, for each module, the "query" is the concatenation of all projected features (F_U), while the "key" and "value" are the individual projected features.

After CrossAttention, the results of the individual attention modules are summed (F_S) to integrate the

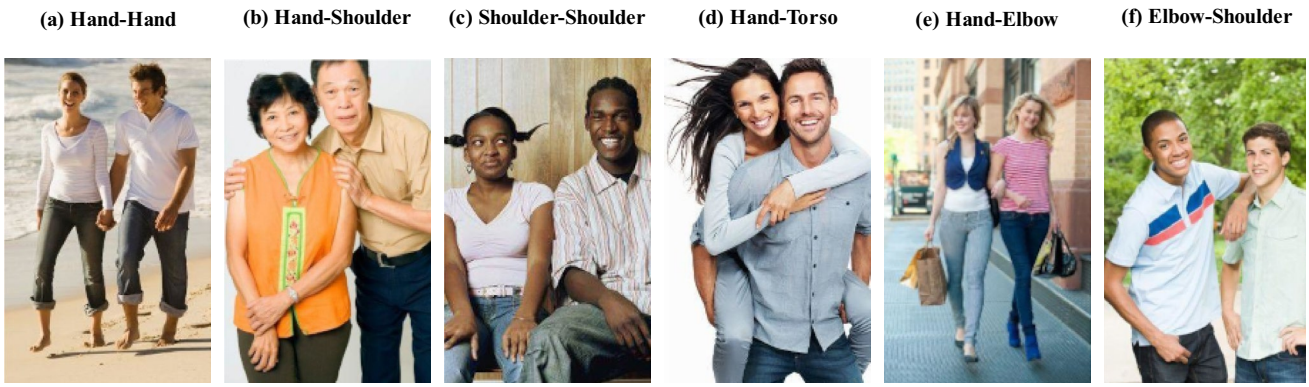

Fig. 5 Touch codes in Proxemics. Images showing the six specific "touch codes" in the Proxemics Dataset

information from the different branches. This aggregation allows combining the essential features of each branch. The F_S result is then fed into a SelfAttention layer (SA), which processes the sum of the attention modules, allowing the model to refine and improve its internal interpretation of the data.

After the SelfAttention, a residual connection is made by summing the output of the Self-Attention layer with its original inputs (F_S), thus generating "F_{CAF} tokens" (Contextual Attention Features). This helps preserve the original information between the model's layers and adds refinements from the SelfAttention, improving training stability. Lastly, the F_{CAF} tokens are averaged and fed into a classification layer.

Finally, since the first task to be addressed is a multi-label problem where each pair can be classified with more than one type of proxemics, the output layer is a 6-unit Sigmoid layer (one for each class). However, in the case of the second task, the problem is multi-class since a pair can only belong to one type of social relation. Thus, the output layer is a 6-unit Softmax layer in this case.

4 Experimental setup

In this section, we will present the two datasets we have used for training our models (Sect. 4.1) and explain the metrics used for the evaluation of these models (Sect. 4.2). Finally, we will see the implementation details such as the preprocessing performed on the images, the partitions realized on the dataset samples, and the training details (Sect. 4.3).

4.1 Datasets

4.1.1 Proxemics

The Proxemics dataset [29] is an annotated database with body joint positions and "touch code" labels, introduced in [2]. This dataset comprises 1,178 images, with 589 being unique images and the other 589 being mirror-flipped versions of the original images.

These images consist of personal photos of family and friends collected from web searches on Flickr, Getty Images, and image search engines on Google and Bing. All images are in color (RGB) and contain two to six people (most images usually have only two) in various poses, scene arrangements, and scales.

For every image in this collection, the dataset creators labeled the positions (coordinates) of the ten major body joints of all the individuals present. These body joints include the head, neck, right and left shoulders, right and left elbows, right and left wrists, and right and left hands. They also labeled all types of proximity between pairs, which are denoted as Hand-Hand (HH), Hand-Shoulder (HS), Shoulder-Shoulder (SS), Hand-Torso (HT), Hand-Elbow (HE), and Elbow-Shoulder (ES) (see Fig. 5).

Finally, after analyzing all the proxemics labeled within the dataset, the researchers observed that most pairs of individuals exhibited either zero (indicating the absence of any of the six types of proxemics) or one "touch code". However, many images also displayed two or more "touch codes." This could occur, for example, when a single person's arm, including the elbow and hand, makes contact with another person's body.

As in the previous work [13], we have approached the problem of proxemics classification at the pair level instead of at the image level as did the authors of the dataset [2]. In

Fig. 6 Types of fine social relationships in PISC. Images showing the six types of fine relationships in the People in Social Context (PISC) Dataset

this way, we have the different types of proxemics appearing in each pair that we can find in the original images.

4.1.2 People in social context–PISC

The People in Social Context (PISC) dataset [7] is a dataset that focuses on social relationships. It consists of 22,670 annotated images with nine types of social relationships. All images are in color (RGB) and present an average of 3.08 persons per image.

In particular, the authors propose two different types of annotations. The first focuses on social domains, splitting the dataset into three types of coarse relationships: No Relation, Intimate Relation, and Non-Intimate Relation. The second considers fine-grained social relationships, represented by six classes: Friends, Family Members, Couples, Professional, Commercial, and No Relation.

Specifically, they provide the bounding box annotations of each person, image information (image source, image size, image ID), and the social relationship annotations (coarse and fine) of all pairs of people in the images.

In this work, we have only experimented with the fine-grained annotations (see Fig. 6) as we have considered them more challenging than the other cases.

4.2 Metrics

We have used the Average Precision (AP) metric to evaluate the proposed variants of our Proxemics-Net++ model.

Average Precision (AP) is a metric designed to evaluate the performance of a model within each class rather than providing a generalized assessment. This approach allows us to assess whether the model exhibits consistent behavior across all classes, indicating that it is a balanced model. On the other hand, a high degree of variability in AP values among different classes might suggest that the model excels in classifying specific classes while struggling with others.

The closer the value of this metric is to 1, the better our model performs within each class.

Finally, once we have obtained the average precision for each class, we calculate the mean Average Precision (mAP), which is the average of the AP values obtained with the six classes.

4.3 Implementation details

4.3.1 Image preprocessing

To work with both datasets, we have applied the same image preprocessing proposed in our previous work [13]. Specifically, we have made clippings of all the pairs and the individuals that compose it in all the RGB images. To perform the clippings, we have used Detectron2 [30], an object and person detector that returns the bounding boxes of each person present in an image. Thus, we have obtained three clippings for each pair in an image: one for each person in the pair and a third clipping with both members of the pair (see the blue box in Fig. 3).

Since the ConvNeXt model works with 224×224 resolution input images, we have preprocessed the clippings by adding padding to make them square and resizing them to 224×224 resolution without changing the aspect ratio.

In the case of the body pose representation, we also obtained three clippings per pair (one for each individual and one for the pair). However, since when generating our body pose representation from the DensePose output, we only kept the body parts referring to the torso, head, and arms, we centered the cropping of the body pose representation on this area of the body (see green box in Fig. 3).

Subsequently, as in the case of the RGB images, we have preprocessed the clippings by adding padding to make them square and resizing them to 224×224 resolution.

4.3.2 Dataset partitions

We have used the same train, test, and validation partitions proposed by the authors of the datasets.

In the case of the Proxemics dataset [2], the authors only propose the train and test partitions, so we have taken approximately 10% of the samples from the training set to use for validation.

In the case of the PISC dataset [7], the train, test, and validation sets have 49,017, 15,497, and 14,536 images.

In particular, a two-fold cross-validation method (set1/set2) has been employed in all our models, as was done by the authors in [2, 7]. In this way, we can directly compare our results with those of the reference papers under the same experimental conditions and see how the model performs on average.

In addition, during training and before merging the information from the different branches, we applied data augmentation techniques to each branch to train our models with a wider variety of samples. Specifically, for RGB images, our augmentation strategies included horizontal flipping, zooming and brightness adjustment. In contrast, for the body pose representation, data augmentation just included horizontal flipping. It is important to note that when horizontal flipping is applied to the RGB images, the same flip is simultaneously applied to the corresponding body pose representations to ensure alignment and consistency.

4.3.3 Training details

As in the previous work, we have used the ConvNeXt architecture as the backbone of our model since it was the one that obtained the best results. Specifically, we have selected two pre-trained ConvNeXt models (without freezing weights) of different complexity, a *"Base"* model [3] and a *"Large"* model [4].

Thus, we will have two variants of our Proxemics-Net++ model (Base and Large). It should be noted that pre-trained models are not combined between branches within the same variant.

In order to train and obtain the best results, we have tested different batch sizes (6, 12, and 18) and used the Adam optimizers. Regarding the learning rate, we have varied between $1 \cdot 10^{-2}$, $1 \cdot 10^{-3}$, $1 \cdot 10^{-4}$, $5 \cdot 10^{-5}$. It is worth noting that in all training, we use the Keras function "keras.callbacks.ReduceLROnPlateau" [5] to automatically adjust the learning rate of the model in case the validation results do not improve in successive epochs. Specifically, we reduce the learning rate if no improvement in validation is observed after six consecutive epochs. Finally, we have selected *binary_crossentropy* as the loss function.

5 Experimental results

In this section, we will show and discuss the best results obtained for the three types of Proxemics-Net++ models we propose (RGB, Pose, and RGB+Pose) in their two variants (Base and Large) and with the two types of fusion implemented (Concatenation and CrossAttention). In addition, we will compare the current state of the art of the two tasks we want to address in this paper (proxemics and social interactions). First, we will show the results obtained on the Proxemics dataset (Sect. 5.1) and then on the PISC dataset (Sect. 5.2).

5.1 Results on the Proxemics dataset

Table 2 shows the best results obtained on the Proxemics dataset for each of the three types of models proposed with Proxemics-Net++: i) RGB models using only the RGB information of the images (first two rows), ii) Pose models using as input only the body pose representation obtained from DensePose (rows three and four) and iii) RGB+Pose models combining both RGB and Pose information (last four rows).

In addition, we also show, for each of these three models, the results obtained when we use the full model (the three branches in the case of the RGB and Pose models or the six branches in the case of the RGB+Pose model) versus when we disable the individual branches (RGB-Individuals and Pose-Individuals). In this way, we can analyze how the additional information of the individuals of a pair contributes to each type of model. These results are presented for both variants of Proxemics-Net++: Base and Large, with each Fusion Block: Concatenation and CrossAttention.

First, we focus on the results obtained with the RGB model, highlighting that those corresponding to the Concatenation fusion were already obtained and shown in our previous work [13]. Upon analysing the model inputs, it is observed that in both Fusion Blocks, the %mAP results for both variants show a significant improvement when incorporating the RGB information of the individual members (second row). This underlines the importance of the RGB information of each pair member (RGB-Individuals branches) in the context of proxemic recognition problem, as it provides additional details about the members of a pair, allowing the model to focus more effectively on task-relevant aspects. On the other hand, when evaluating fusion type, the CrossAttention Fusion Block outperforms Concatenation, showing a decrease in standard deviation and a notable increase from

[3] The pre-trained *Base* model is located in: https://dl.fbaipublicfiles.com/convnext/convnext_base_22k_224.pth

[4] The pre-trained *Large* model is located in: https://dl.fbaipublicfiles.com/convnext/convnext_large_22k_224.pth

[5] keras.callbacks.ReduceLROnPlateau function: https://keras.io/api/callbacks/reduce_lr_on_plateau/

Table 2 Best results obtained on the Proxemics dataset for each of the three proposed model types, using the two variants of Proxemics-Net++ and the two proposed Fusion Blocks

Model				Concatenation mAP(Set1-Set2)		CrossAttention mAP(Set1-Set2)	
RGB pair	RGB individuals	Pose pair	Pose individuals	ConvNeXt base	ConvNeXt large	ConvNeXt base	ConvNeXt large
✓				62.3 ± 3.3	61.5 ± 2.5	66.1 ± 1.5	63.6 ± 0.2
✓	✓			63.3 ± 2.4	63.9 ± 2.9	67.5 ± 2.1	68.4 ± 1.4
		✓		67.7 ± 3.2	64.4 ± 4.4	69.8 ± 1.2	68.4 ± 1.4
		✓	✓	64.8 ± 0.7	**67.1 ± 4.1**	69.4 ± 2.0	**70.8 ± 4.7**
✓		✓		**70.0 ± 1.5**	65.5 ± 3.5	71.3 ± 4.2	68.5 ± 3.2
✓	✓	✓		68.5 ± 3.4	65.0 ± 3.6	70.2 ± 2.1	66.7 ± 2.5
✓		✓	✓	68.4 ± 3.5	64.4 ± 2.2	71.5 ± 3.7	69.4 ± 2.8
✓	✓	✓	✓	67.8 ± 2.4	64.3 ± 2.1	**71.6 ± 1.7**	69.9 ± 2.2

Best results of %mAP (together with their standard deviation) obtained when using only RGB information (first two rows), pose information (rows three and four) or the combination of RGB+Pose (last four rows). In addition, for each of these three models, we show the results obtained when we disable the individual branches (RGB-Individuals or Pose-individuals). These results are shown for the two variants of Proxemics-Net++ (Base and Large) with the two proposed Fusion Blocks: Concatenation and CrossAttention

Bold values represent the best result obtained in each of the four variants of Proxemics-Net++ (Concatenation with Base and Large and CrossAttention with Base and Large)

63.3 ± 2.4 to 67.5 ± 2.1 in the Base variant and from 63.9 ± 2.9 to 68.4 ± 1.4 in the Large variant, reflecting greater consistency and effectiveness of the model with CrossAttention.

In the Pose model, an increase in %mAP is observed for all cases compared to the RGB model. This indicates that the body pose representation provides relevant information for proxemic recognition in images, achieving good results without the need to incorporate RGB information. It is worth noting in the Pose model, unlike the RGB model, the inclusion of the pose information of the individuals of the pair (fourth row) only improves the results in the Large variant (for both Fusion Blocks), being the model with the best results for this variant among the three types of models proposed. Once again, an improvement in results with CrossAttention fusion compared to Concatenation is observed in both variants.

In the RGB+Pose model, we observe that the use of the Base variant as backbone in both Fusion Blocks produces the highest values of %mAP in all cases compared to the other two proposed models. In particular, in the Concatenation fusion, the RGB+Pose model with only the RGB-Pair and Pose-Pair branches active (row five) achieves the best results with 70.0 ± 1.5 mAP. In the CrossAttention fusion, the RGB+Pose model with all active branches (last row) shows the highest mAP with 71.6 ± 1.7 and the lowest standard deviation. In the case of the Large variant, in both Fusion Blocks, all cases of the RGB+Pose model outperform the RGB model but but not the Pose model. However, it is worth noting that in the CrossAttention fusion, the results of RGB+Pose are much closer to those of the Pose model. For example, the RGB+Pose model with all active branches (last row) achieves a mAP of 69.9 ± 2.2 compared to 70.8 ± 4.7 for the Pose model (row four). Once again, in both variants, an improvement in results is observed with the CrossAttention fusion compared to the Concatenation, highlighting the effectiveness of the CrossAttention fusion in integrating RGB and Pose information in the RGB+Pose model.

Thus, as a summary of this comparison, we can state that the best results have been obtained in all cases with the CrossAttention fusion, demonstrating its effectiveness in the problem of proxemic recognition. For the Base variant, the RGB+Pose model incorporating all branches (pairs and individuals) achieved the highest results with a 71.6 ± 1.7 % mAP, while in the Large variant, the best performance was observed in the Pose model, which includes both Pose-Individuals and Pose-Pair branches, achieving a 70.8 ± 4.7 % mAP. This indicates that the inclusion of body pose information helps significantly in the problem of proxemic recognition in images, as its incorporation allows us to improve the results considerably with respect to only using RGB. This makes sense since we are classifying proxemics or "touch-codes" in which there is always physical contact, and the body pose plays an important role.

5.1.1 Comparison to the state of the art

Table 3 compares our best model with the existing state-of-the-art in the proxemics recognition problem. Since the existing works addressing the problem of proxemic recognition in images are not very recent, with the exception

Table 3 Comparison of our best model obtained on the Proxemics dataset with the state of the art

Model	HH	HS	SS	HT	HE	ES	mAP (a)	mAP (b)
Yang et al. [2]	37	29	50	61	38	34	42	38
Chu et al. [16]	41.2	35.4	62.2	–	43.9	55	–	46.6
Jiang et al. [17]	59.7	52	53.9	33.2	36.1	36.2	45.2	47.5
Li W. et al. [20]*	56.7	55.1	52.8	78.4	65.0	65.5	62.3	59.1
Sousa et al. [23]*	66.2	55.1	69.5	78.8	65.6	68.1	67.2	64.9
Jiménez et al. [13]	62.4	56.7	62.4	**86.4**	68.8	67.9	67.4	63.8
Our ConvNeXt_Base (Cross)Model	HH	HS	SS	HT	HE	ES	mAP (a)	mAP (b)
(RGB+Pose–full model)	**71.5**	**63.2**	**80.5**	80.7	**75.6**	**71.3**	**73.8**	**72.4**

The Table shows the average precision (%) in proxemic recognition: mAP(a) is the average of all classes (Set1 and Set2), and mAP(b) excludes the HT class. *These results have been computed by adapting the code released by the authors of the methods.

Bold values represent the best result obtained in each of the six labels and in mAP (among allthe models compared)

of our previous work [13], we update our comparison by incorporating two more recent methods that do not directly address the proxemic recognition problem but do address the Human Interaction Recognition problem. Specifically, we have selected the works [20, 23], which consist of methods applied to the PISC dataset and which are considered the current state of the art for that dataset, as reflected in the results presented in Table 5. For this experimental study, we adapted the Proxemics dataset to the input specifications required by these methods and slightly modified the methods to handle a multi-label classification problem, as the PISC dataset presents a multi-class problem. This change included adjusting the output layer of the models and the way labels were handled so that the new methods could work efficiently with multiple labels simultaneously, thus adapting to the specific challenge presented by Proxemics. Once these changes were made, a series of experiments were conducted to evaluate the performance of both methods. The best results obtained for both methods are detailed in Table 3 (see fourth and fifth rows). In this way, we not only compare our proposed method with the existing state of the art on Proxemics, but also evaluate its effectiveness against more recent methods, following its application in the context of Proxemics.

In this Table, two values of %mAP are compared: mAP(a) is the value of mAP explained in the previous sections (the mean of the AP values of the six types of proxemics), and mAP(b) is the mean of the AP values but excluding the Hand-Torso (HT) class as done in [16].

Since our %mAP results are obtained at the pair level rather than at the image level, we had to re-evaluate our best model obtained in the previous Table (see Table 2) to compare our results with the state of the art. For each image in the Proxemics dataset, the Proxemics-Net++ network processes all possible pairs. The image-level result is calculated as the maximum classification score obtained among all image pairs in each proxemics class.

Looking at Table 3, we observe that our best model (RGB+Pose with all individual and pair branches activated (full model), with Base variant and CrossAttention Fusion Block) obtains the best %mAP results in almost all types of proxemics to be classified and in both comparisons (mAP(a-b)). Specifically, 73.8% vs. 67.4% of mAP(a) and 72.4% vs. 64.9% of mAP(b)). Thus, we outperform the current state of the art by a significant margin, with improvements of up to 6.4% for mAP(a) and 7.5% for mAP(b).

Therefore, these results show that RGB information combined with body pose information and a state-of-the-art deep learning model such as ConvNeXt does help in the problem of proxemics recognition in images (the first task to be addressed in this paper) since it considerably improves the results obtained in the previous work [13] and by all competing models.

5.1.2 Failure cases on the Proxemics dataset

Figure 7 shows some images, together with their corresponding pose estimation, that have been misclassified by our best model. Based on our observation, the pose estimation in all three images seems inaccurate. This is because some body parts, such as arms, have not been correctly detected. As a result, our model may have classified the different types of proxemics in these images incorrectly. The poor pose estimation may be due to the characteristics of the images since, as we can observe, the three images show pairs of people overlapping each other, with very similar clothing colors and in which there are occluded body parts, which may have

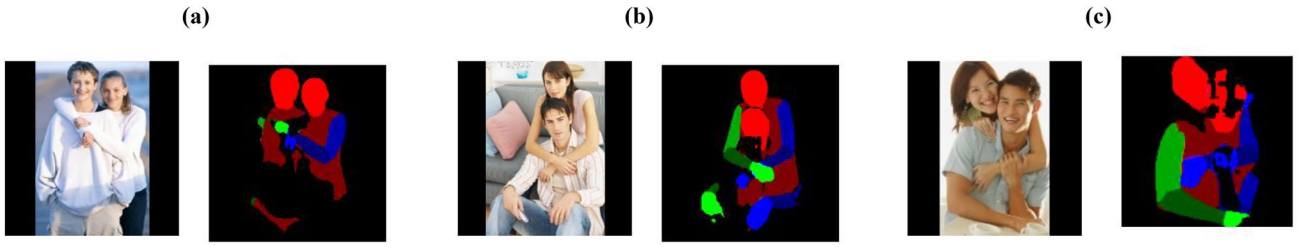

Fig. 7 Failure cases on the Proxemics dataset. Images classified incorrectly by our model show a bad estimation of the pose. All the images present a great occlusion of the body parts and couple with similar clothing colors that could have caused a worse pose estimation (color figure online)

Table 4 Best results obtained on the PISC dataset for each of the three proposed model types, using the two variants of Proxemics-Net++ and the two proposed Fusion Blocks

Model				Concatenation mAP(Set1-Set2)		CrossAttention mAP(Set1-Set2)	
RGB pair	RGB individuals	Pose pair	Pose individuals	ConvNeXt base	ConvNeXt large	ConvNeXt base	ConvNeXt large
✓				58.4 ± 1.6	57.2 ± 1.3	64.4 ± 1.2	57.9 ± 1.3
✓	✓			**70.1 ± 2.9**	**64.7 ± 1.6**	**69.6 ± 0.6**	**65.0 ± 0.7**
		✓		43.3 ± 1.1	27.9 ± 0.5	51.2 ± 0.6	45.1 ± 0.2
		✓	✓	42.0 ± 1.3	44.2 ± 0.1	50.9 ± 0.6	46.3 ± 2.1
✓		✓		55.3 ± 1.8	55.6 ± 2.5	56.6 ± 2.1	61.7 ± 1.1
✓	✓	✓		61.8 ± 1.4	51.6 ± 0.6	63.9 ± 1.1	57.7 ± 0.4
✓		✓	✓	57.5 ± 1.0	51.6 ± 0.5	62.6 ± 2.6	61.1 ± 2.2
✓	✓	✓	✓	55.1 ± 2.7	54.6 ± 0.4	55.4 ± 1.9	57.7 ± 0.6

Best results of %mAP (together with their standard deviation) obtained when using only RGB information (first two rows), pose information (rows three and four) or the combination of RGB+Pose (last four rows). In addition, for each of these three models, we show the results obtained when we disable the individual branches (RGB-Individuals or Pose-individuals). These results are shown for the two variants of Proxemics-Net++ (Base and Large) with the two proposed Fusion Blocks: Concatenation and CrossAttention

Bold values represent the best result obtained in each of the four variants of Proxemics-Net++ (Concatenation with Base and Large and CrossAttention with Base and Large)

confused the DensePose pose estimator in distinguishing each individual in the image.

Therefore, the performance of our model depends on the quality of the pose estimation. Future pose estimators may be able to solve this problem even better when there is excessive occlusion and confusion of the pairs.

5.2 Results on the PISC dataset

Table 4 shows the best results obtained on the PISC dataset for each of the three types of models proposed with Proxemics-Net++: i) RGB models using only the RGB information of the images (first two rows), ii) Pose models using as input only the body pose representation obtained from DensePose (rows three and four) and iii) RGB+Pose models combining both RGB and Pose information (last four rows).

In addition, we also show, for each of these three models, the results obtained when we use the full model (the three branches in the case of the RGB and Pose models or the six branches in the case of the RGB+Pose model) versus when we disable the individual branches (RGB-Individuals and Pose-Individuals). In this way, we can analyze how the additional information of the individuals of a pair contributes to each type of model. These results are presented for both variants of Proxemics-Net++: Base and Large, with each Fusion Block: Concatenation and CrossAttention.

Focusing on the RGB model results, we notice that %mAP results for both fusion types improve when we include the RGB information of the individual member (second row). This suggests that RGB information from both individuals in a pair significantly contributes to the social relation recognition problem in images, as opposed to using only the RGB information of the pair. Additionally, comparing fusion types, the CrossAttention Fusion Block often yields similar

Table 5 Comparison of our best model obtained on the PISC dataset with the state of the art. The Table shows the average accuracy (%AP) of each type of social interaction recognition as well as the average of all classes (%mAP)

Model	Friends	Family	Couple	Prof.	Comm.	No rel.	mAP
Li J. et al. [7]	60.6	64.9	54.7	82.2	58	70.6	65.2
Zhang et al. [18]	64.6	67.8	60.5	76.8	34.7	70.4	70.0
Goel et al. [19]	–	–	–	–	–	–	71.6
Li W. et al. [20]	60.8	65.9	**84.8**	73.0	51.7	70.4	72.7
Li L. et al. [21]	**82.2**	39.4	33.2	60.0	47.7	71.8	73.3
Yang et al. [22]	63.1	73.5	78.3	**82.7**	**76.8**	71.8	73.6
Sousa et al. [23]	49.4	70.5	74.6	76.5	59.6	74.6	**75.2**
Our ConvNeXt_ Base (Concat) (RGB–full model)	56.2	**83.9**	77.6	61.0	59.0	**82.9**	70.1

Bold values represent the best result obtained in each of the six labels and in mAP (among allthe models compared)

or better %mAP results, but with a notably reduced standard deviation, like in the RGB–Full model (second row). This reflects greater consistency and effectiveness with the CrossAttention model.

In the Pose model, the results for both fusion types are inferior to those of the RGB model. This may indicate that our representation of the body pose (designed more for the proxemics problem and focused more on the detection of the different physical interactions) does not provide relevant information in the problem of the recognition of the type of social relationships in images, in which there may or may not be physical interaction. Notably, unlike the RGB model, including individual body pose information in the Pose model (fourth row) only enhances results in the Large variant for both fusion types. A notable improvement with CrossAttention over Concatenation is observed in both variants, as it consistently yields better %mAP results with lower standard deviation.

In the RGB+Pose model, %mAP results for both fusion types and variants show improvement over the Pose model but are still not as good as the RGB model. The combination of RGB and Pose information from pairs (RGB-Pair and Pose-Pair), along with individual RGB data (RGB-Individuals) and CrossAttention fusion, performs best in the RGB+Pose model with a 63.9 ± 1.1% mAP (sixth row, Base variant). This suggests that appearance information is more significant for this problem. Again, in both variants, an improvement in results with CrossAttention over Concatenation is observed, highlighting the effectiveness of CrossAttention in integrating RGB and Pose information in the RGB+Pose model.

In summary, across both variants and fusion types, the full RGB model incorporating both individual (RGB-Individuals) and pair (RGB-Pairs) branches achieves the best results. Specifically, the Base variant with Concatenation fusion performs the best, achieving a 70.1 ± 2.9% mAP. It is important to highlight that while the RGB model with Concatenation fusion achieves the highest %mAP, the same model with CrossAttention fusion (second row) could also be considered as optimal as it has a similar %mAP result but a lower standard deviation (70.1 ± 2.9% mAP vs. 69.6 ± 0.6% mAP), indicating similar performance. Except for this case, CrossAttention fusion has significantly improved all proposed models in both variants with respect to Concatenation fusion, demonstrating its effectiveness in social relation recognition.

Given that in both cases, the incorporation of pose information has not improved the results, it could be stated that such information does not help recognize types of social relations in images. This may be due to several factors: (1) we are dealing with a problem in which there may or may not be physical interaction between pairs of people, so perhaps the pose is not a feature that contributes in the same way as in proxemics; (2) the same pose may appear in two different social relationships, for example, a boy eating with a friend or with a brother; and, (3) the mislabeling of the images. As we will see in Subsubsection 5.2.2, the dataset has mislabeled images both at the level of the type of interaction (they are very ambiguous) and at the level of labeling of the individuals of the couples since there are people labeled with only a small part of the body that is difficult to detect with the human eye, which has generated worse estimates of the pose in certain images and therefore, worse results when generalizing our models.

5.2.1 Comparison to the state of the art

Table 5 compares our best model with the existing state-of-the-art in the social interaction recognition problem.

Looking at the Table, we can see that our best model (RGB model incorporating both individual and pairs branches with Base variant and Concatenation Fusion Block) obtains the best results in the *Family* and *No Relation* categories but does not outperform on average the current state of the art (70.1% of mAP vs. 75.2% of mAP).

Fig. 8 Failure cases on the PISC dataset. Failure cases encountered with the PISC dataset. **a** Images with ambiguous labeling and incorrectly classified by our best model. The green rectangle shows the ground truth, and the red rectangle shows the incorrect prediction of our model. **b** Images in which people are difficult to recognize, or only parts of people's bodies are labeled

It should be noted that all the works shown in the Table, except for the work of the authors of the PISC dataset [7] (first row), use graph-based architectures to solve this problem, which is a significant difference between our model and those of the state of the art.

If we compare with [7], whose work does use a deep neural network, we can see that our model obtains better results using only the RGB information together with the ConvNeXt architecture as the backbone (70.1% of mAP vs. 65.2% of mAP). Therefore, comparing with similar architectures, we can affirm that the RGB information, unlike the pose information, can help significantly in recognizing social interactions in images.

Even so, seeing that in recent years, the problem of recognition of social interactions in images has been oriented more towards graph-based architectures, we could deduce that it is a problem that, by its nature and ambiguity, needs to be treated with another type of architecture more focused on the relationships between nodes, features, etc.

5.2.2 Failure cases on the PISC dataset

The following are some failures encountered with respect to the problem of social relation recognition, which may have led to worse results, especially when incorporating pose information.

Ambiguous labeling of numerous images After evaluating our best model (see Sect. 5.2.1) we encountered some really ambiguous false positives. Figure 8a shows two examples of images misclassified by our model. The green rectangle shows the ground truth, and the red rectangle shows the incorrect prediction of our model. As we can see in the image on the left, the image is labeled as a friendship relationship, but with visual information alone, it is impossible to be 100% sure that it is a friendship relationship rather than a family relationship, for example. In the image on the right, something similar happens. The actual relationship is professional, but it could be family or friendship at first glance. This shows that the dataset has ambiguously labeled images that can lead to a worse generalization of our models.

Labeling of individuals who are not very visible or relevant When we applied the DensePose estimator to the images of the PISC dataset, we found some images in which this estimator only estimated body parts of the people who were actually labeled in the dataset or even was not able to detect as many people as were labeled (in this second case, we had to lower the detection threshold of the DensePose estimates so that it would detect such people even though it returned worse results). In Fig. 8b, we show some images in which people are labeled in the dataset that are difficult to see because of their large occlusion or irrelevance in the image. Consequently, obtaining poor pose estimates in certain images due to mislabeling in some images may have caused us to obtain results showing that the pose information does not help in this problem when it may have.

Furthermore, since the PISC dataset includes ground-truth bounding boxes (BBs) for all individuals in the images, we decided to evaluate the impact of using these perfect clippings of all individuals versus the clippings obtained using the object and person estimator, Detectron2 [30], in a dataset where person recognition is challenging. To this end, we trained a new model using the best previously obtained configuration (RGB model—full model with Concatenation fusion, see Table 5) and ground-truth clippings for all individuals. We observed a significant improvement, with a %mAP (set1-set2) of 77.2 ± 2.5, compared to 70.1 ± 2.9 obtained using Detectron2 clippings. This result not only outperforms our best model, but also improves on the current state of the art, which stands at 75.2%.

The enhanced performance when using ground-truth BBs compared to those provided by Detectron2 indicates our model's improved capability in recognizing social relationships. This outcome highlights the influence of the quality and accuracy of the clippings generated by

Detectron2 in challenging datasets. Therefore, implementing a more accurate person detector could lead to significant improvements in the model's results.

6 Conclusion and future work

In this work, we have proposed Proxemics-Net++, an extension of the Proxemics-Net model [13], capable of addressing the challenge of detecting human interactions in images from two different tasks and using both RGB and Pose information. One task, from a more detailed perspective, focused on detecting and classifying the types of physical interactions between people, known as proxemics, and the other, from a broader perspective, focused on detecting and classifying different social interactions between pairs of individuals in which there may or may not be physical interactions.

Proxemics-Net++ has six inputs: three for RGB information of the couple and the individuals that compose it, plus three inputs that have been added for representing the body pose of both the couple and the individuals. Additionally, this model has been evaluated using the state-of-the-art deep architecture ConvNeXt as the backbone, specifically with two of its variants, Base and Large. In addition, we have implemented a CrossAttention fusion, which has improved the performance of our model compared to simple Concatenation fusion.

In the case of the first proposed task, our results on the Proxemics dataset have demonstrated that body pose information, combined with RGB information and state-of-the-art deep architecture such as ConvNeXt, significantly contributes to the problem of proxemics recognition in images since, using it, we outperform the existing state of the art. Regarding the second task, our experiments on the PISC dataset show that due to the nature of the problem and the dataset's ambiguity, body pose information does not contribute as significantly to this task, with the best results obtained using only RGB information. In addition, the Base variant has been the best backbone in both tasks.

Future work includes testing Proxemics-Net++ with the other type of annotation provided by the authors of PISC, which focuses on the social domain (no relation, intimate, and non-intimate relation). This will help analyze whether body pose information provides relevant insights into the problem of classifying social relationships at a more general level and not as fine-grained and ambiguous as we have seen in cases where a couple could belong to multiple categories (friends and family). In this context, the pose and proximity of individuals can help determine whether the relationship is intimate. We also plan to test our model with alternative representations of the pose that include the lower body, for example, and using other architectures as backbone. Furthermore, since we have observed a trend in using graph-based architectures for the PISC dataset, another avenue for future work will be to test our pose representation with these architectures to determine if we can achieve improved results.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by [Isabel Jiménez-Velasco]. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Funding for open access publishing: Universidad de Córdoba/CBUA. Supported by the MCIN Project TED2021-129151B-I00/ AEI/10.13039/ 501100011033/European Union NextGenerationEU/ PRTR, and project PID2019-103871GB-I00 of the Spanish Ministry of Economy, Industry and Competitiveness, FEDER.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethical approval For this article, the authors did not undertake work that involved humans or animals.

References

1. Patron A, Reid I, Marszalek M, Zisserman A (2012) Structured learning of human interactions in TV shows. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2012.24
2. Yang Y, Baker S, Kannan A, Ramanan D (2012) Recognizing proxemics in personal photos. In: IEEE conference on CVPR. https://doi.org/10.1109/CVPR.2012.6248095
3. Muhamada AW, Mohammed AA (2021) Review on recent computer vision methods for human action recognition. ADCAIJ 10(4):361–379. https://doi.org/10.14201/ADCAIJ2021104361379
4. Le VT, Tran K, Truong V (2022) A comprehensive review of recent deep learning techniques for human activity recognition. Comput Intell Neurosci. https://doi.org/10.1155/2022/8323962
5. Ilyas CMA, Rehm M, Nasrollahi K (2022) Deep transfer learning in human-robot interaction for cognitive and physical rehabilitation purposes. Pattern Anal Appl 25:653–677. https://doi.org/10.1007/s10044-021-00988-8
6. Gutoski M, Lazzaretti AE, Lopes HS (2023) Unsupervised open-world human action recognition. Pattern Anal Appl. https://doi.org/10.1007/s10044-023-01202-7

7. Li J, Wong Y, Zhao Q (2020) Visual social relationship recognition. IJCV 128:1750–1764. https://doi.org/10.1007/s11263-020-01295-1
8. Tanisik G, Zalluhoglu C, Ikizler N (2021) Multi-stream pose convolutional neural networks for human interaction recognition in images. Signal Process Image Commun 95:116265. https://doi.org/10.1016/j.image.2021.116265
9. Lee DG, Lee SW (2022) Human interaction recognition framework based on interacting body part attention. Pattern Recognit. 128:108645. https://doi.org/10.1016/j.patcog.2022.108645
10. Sun R, Zhang Q, Luo C et al (2022) Human action recognition using a convolutional neural network based on skeleton heatmaps from two-stage pose estimation. Biomim Intell Robot 2:100062. https://doi.org/10.1016/j.birob.2022.100062
11. Dosovitskiy A, Beyer L, Kolesnikov A, et al (2021) An image is worth 16x16 words: transformers for image recognition at scale. ICLR
12. Liu Z, Mao H, Wu CY (2022) A convnet for the 2020s. In: IEEE/CVF conference on CVPR. https://doi.org/10.1109/CVPR52688.2022.01167
13. Jiménez I, Muñoz R, Marín MJ (2023) Proxemics-net: automatic proxemics recognition in images. In: Iberian conference on pattern recognition and image analysis, pp 402–413. https://doi.org/10.1007/978-3-031-36616-1_32. IbPRIA 2023
14. Guler RA, Neverova N, Kokkinos L (2018) Densepose: dense human pose estimation in the wild. In: Proceedings of the IEEE conference on CVPR, pp 7297–7306.https://doi.org/10.1109/CVPR.2018.00762
15. Edward TH (1963) A system for the notation of proxemic behavior. Am Anthropol 65(5):1003–1026. https://doi.org/10.1525/aa.1963.65.5.02a00020
16. Chu X, Ouyang W, Yang W (2015) Multi-task recurrent neural network for immediacy prediction. In: Proceedings of the IEEE international conference on computer vision, pp 3352–3360. https://doi.org/10.1109/ICCV.2015.383
17. Jiang H, Grauman K (2017) Detangling people: individuating multiple close people and their body parts via region assembly. In: IEEE conference on CVPR, pp 3435–3443. https://doi.org/10.1109/CVPR.2017.366
18. Zhang M, Liu X, Liu W (2019) Multi-granularity reasoning for social relation recognition from images. In: IEEE international conference on multimedia and expo (ICME), pp 1618–1623. https://doi.org/10.1109/ICME.2019.00279
19. Goel A, Ma K, Tan C (2019) An end-to-end network for generating social relationship graphs. In: IEEE/CVF conference on CVPR, pp 11178–11187. https://doi.org/10.1109/CVPR.2019.01144
20. Li W, Duan Y, Lu J (2020) Graph-based social relation reasoning. In: European conference on computer vision, pp 18–34. https://doi.org/10.1007/978-3-030-58555-6_2
21. Li L, Qing L, Wang Y (2022) HF-SRGR: a new hybrid feature-driven social relation graph reasoning model. Vis Comput 38:3979–3992. https://doi.org/10.1007/s00371-021-02244-w
22. Yang X, Xu F, Wu K (2021) Gaze-aware graph convolutional network for social relation recognition. IEEE Access 9:99398–99408. https://doi.org/10.1109/ACCESS.2021.3096553
23. Sousa EV, Macharet DG (2023) Structural reasoning for image-based social relation recognition. Comput Vis Image Underst 235:103785. https://doi.org/10.1016/j.cviu.2023.103785
24. Farrajota M, Rodrigues JMF, Du JMH (2019) Human action recognition in videos with articulated pose information by deep networks. Pattern Anal Appl 22:1307–1318. https://doi.org/10.1007/s10044-018-0727-y
25. Bertoni L, Kreiss S, Alahi A (2021) Perceiving humans: from monocular 3d localization to social distancing. IEEE Trans Intell Transp Syst. https://doi.org/10.1109/TITS.2021.3069376
26. Russakovsky O, Deng J, Su H (2015) ImageNet large scale visual recognition challenge. IJCV 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y
27. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF ICCV
28. Liu Z, Courant R, Kalogeiton V (2023) Funnynet: audiovisual learning of funny moments in videos. In: Computer vision—ACCV 2022, pp 433–450. https://doi.org/10.1007/978-3-031-26316-3_26
29. Yang Y, Baker S, Kannan A, Ramanan L (2012) PROXEMICS dataset. https://www.dropbox.com/s/5zarkyny7ywc2fv/PROXEMICS.zip?dl=0. Last visited: 26-October-2023
30. Wu Y, Kirillov A, Massa F, et al (2019) Detectron2. https://github.com/facebookresearch/detectron2. Last visited: 26-October-2023