UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Eduardo Vieira e Sousa

Raciocínio Estrutural Multiescala para Reconhecimento de Relações Sociais em Imagens

Belo Horizonte
2021

Eduardo Vieira e Sousa

Raciocínio Estrutural Multiescala para Reconhecimento de Relações Sociais em Imagens

Versão final

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientador: Douglas Guimarães Macharet

Belo Horizonte
2021

Eduardo Vieira e Sousa

Structure-Aware Multi-Scale Reasoning for Image-based Social Relation Recognition

Final version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Douglas Guimarães Macharet

Belo Horizonte
2021

Sousa, Eduardo Vieira e

S725s Structure-aware multi-scale reasoning for image-based social relation recognition [manuscrito] / Eduardo Vieira e Sousa — 2021.
xxviii, 103 f. il.

Orientador: Douglas Guimarães Macharet
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação
Referências: f. 93-103

1. Computação – Teses. 2. Visão computacional – Teses. 3. Reconhecimento de padrões – Teses. 4. Aprendizado profundo – Teses. 5. Redes de relações sociais – Teses. 6. Redes neurais de grafos – Teses. I. Macharet, Douglas Guimarães. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. III.Título.

CDU 519.6*82 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6ª Região nº 1510

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

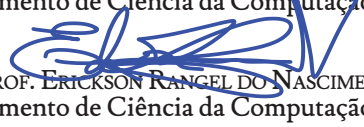
FOLHA DE APROVAÇÃO

Structure-Aware Multi-Scale Reasoning for Image-based Social Relation Recognition

EDUARDO VIEIRA E SOUSA

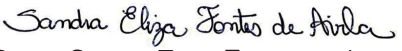
Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. DOUGLAS GUIMARÃES MACHARET - Orientador
Departamento de Ciência da Computação - UFMG



PROF. ERICKSON RANGEL DO NASCIMENTO
Departamento de Ciência da Computação - UFMG

PROF. WILLIAM ROBSON SCHWARTZ
Departamento de Ciência da Computação - UFMG



PROFA. SANDRA ELIZA FONTES DE AVILA
Instituto de Computação - UNICAMP

Belo Horizonte, 29 de Julho de 2021.

Dedico este trabalho à minha irmã, Fernanda. Enquanto eu habitar esse mundo, sua memória viverá em mim.

Acknowledgments

Primeiramente, quero de agradecer aos meus pais, Regina e Fernando, ao meu irmão, Henrique, e à minha namorada, Lívia. Cada um de vocês contribuiu de uma maneira fundamental para que eu pudesse alcançar essa conquista em minha vida.

Sou muito grato ao meu orientador, professor Douglas G. Macharet, pelo apoio durante todo o percurso. Obrigado pela paciência, pelos conselhos e disposição, sempre exercendo seu papel com extrema dedicação e consideração.

Agradeço aos membros da banca, professores Erickson Rangel do Nascimento (DCC/UFMG), William Robson Schwartz (DCC/UFMG) e Sandra Eliza Fontes de Ávila (IC/UNICAMP) pela disponibilidade e pela avaliação cuidadosa do trabalho.

Muito obrigado também aos colegas do VeRLab pelo ambiente receptivo e pelo companheirismo. Por fim, agradeço a todos os excelentes professores do DCC com os quais tive a oportunidade de trabalhar, e que foram essenciais para esse processo de crescimento e aprendizado.

"Far better it is to dare mighty things, to win glorious triumphs, even though checkered by failure, than to take rank with those poor spirits who neither enjoy much nor suffer much, because they live in the gray twilight that knows neither victory nor defeat."

(Theodore Roosevelt)

Resumo

As sociedades modernas são compostas por estruturas complexas que emergem das relações entre indivíduos, e a compreensão desses arranjos tem o potencial de se tornar uma ferramenta poderosa para sistemas inteligentes. Os métodos atuais de reconhecimento de relações sociais baseados em imagens isolam informações específicas da entrada com intuito de capturar aspectos essenciais que definem esses relacionamentos. No entanto, esta é uma abordagem imprecisa para analisar relações sociais, uma vez que a interação entre todas essas partes forma uma estrutura intrincada, sendo tão valiosa quanto as informações que cada componente carrega individualmente. Por esse motivo, é crucial capturar a estrutura social original para alcançar o raciocínio de alto nível necessário para identificar os relacionamentos de forma adequada. Neste trabalho, uma nova abordagem para interpretar métodos de reconhecimento de relacionamento social baseados em imagens é apresentada, considerando três escopos distintos de análise, denominados escalas sociais, relacionados a informações individuais, relativas e gerais. Além disso, também é avaliado como os dados dessas diferentes perspectivas sociais são combinados, levando em conta a capacidade de capturar dependências e restrições em múltiplas escalas. O Social Knowledge Graph (SKG) é proposto com base nas conclusões obtidas da análise conduzida, produzindo uma representação capaz de replicar a estrutura social da imagem de entrada. Essa representação única é explorada por meio da Social Graph Network (SGN), aplicando estratégias específicas de agregação de features, conforme as informações embutidas na estrutura do grafo. O desempenho do método proposto foi avaliado em benchmarks bem estabelecidos, alcançando um novo estado da arte. Finalmente, uma análise profunda da metodologia e dos conceitos por trás dela é conduzida, fornecendo uma visão sobre o processo de decisão do modelo proposto e obtendo resultados que suportam a nova interpretação dos métodos de reconhecimento de relações sociais.

Palavras-chave: Visão Computacional. Reconhecimento de Padrões. Reconhecimento de Relações Sociais. Aprendizado Profundo. Redes Neurais de Grafos.

Abstract

Modern societies are composed of complex structures that emerge from the relationships between individuals, and the comprehension of these arrangements has the potential to become a powerful tool for intelligent systems. Current image-based social relation recognition methods isolate specific information from the input to capture essential aspects defining these relationships. However, this is an inaccurate approach for analyzing social relations, since the interaction between all these parts form an intricate structure, which is as valuable as the information each piece carries individually. For this reason, it is crucial to capture the original social structure to achieve the high-level reasoning required to identify relationships adequately. In this work, a novel approach to interpret image-based social relation recognition methods is presented, considering three distinct scopes named social scales, regarding individual, relative, and general information. Additionally, it also evaluates how the data from these different social perspectives is combined, taking into account the capacity of capturing multi-scale interdependencies and constraints. The Social Knowledge Graph (SKG) is proposed based on the conclusions obtained from the conducted analysis, producing a representation capable of replicating the original social structure from the input image. This unique representation is exploited with the Social Graph Network (SGN) by employing specific feature aggregation strategies according to the information embedded in the graph structure. The performance of the proposed method was evaluated in well-known benchmarks for social relation recognition, achieving a new state-of-the-art. Finally, a deep analysis of the methodology and the main concepts behind it is conducted, providing insight into the decision-making process of the proposed model and delivering results that support the new interpretation of social relationship recognition methods.

Palavras-chave: Computer Vision. Pattern Recognition. Social Relation Recognition. Deep Learning. Graph Neural Networks.

List of Figures

List of Tables

List of Acronyms

CNN	Convolutional Neural Network
ConvGNNs	Convolutional Graph Neural Networks
DSFS	Deep Supervised Feature Selection
GAT	Graph Attention Networks
GCN	Graph Convolutional Network
GGNN	Gated Graph Neural Network
GR^2N	Graph Relational Reasoning Network
GRM	Graph Reasoning Model
GRU	Gated Recurrent Unit
HRI	Human-Robot Interaction
LSTM	Long Short-Term Memory
MGR	Multi-Granularity Reasoning
MPNN	Message Passing Neural Networks
NLNN	Non-Local Neural Networks
NN4G	Neural Network for Graphs
PIPA	People in Photo Album
PISC	People in Social Context
R-CNN	Region-based CNN
RecGNNs	Recurrent Graph Neural Networks
SGD	Stochastic Gradient Descent
SGN	Social Graph Network
SKG	Social Knowledge Graph
SKG+	Social Knowledge Graph Plus
SRG-IN	Social Relationship Graph Inference Network
SRiV	Social Relation in Videos
SSN	Social Scales Network
TSN	Temporal Segment Networks
ViSR	Video-based Social Relation

List of Symbols

ϵ	Constant for numerical stability in the *RelationConv*
Λ	Diagonal matrix of eigenvalues
σ	Sigmoid activation function
Φ	Generic attribute extraction model
$\mathbf{a}$	Attribute node type from the SKG
A	Adjacency matrix representation from a graph
$\mathcal{A}$	Total number of attributes extracted
$\mathbf{activity}(\cdot)$	Activity features extraction model
$\mathbf{age}(\cdot)$	Age features extraction model
B	Set of bounding boxes
$\mathcal{C}$	Total number of relationship classes
$\mathbf{C}(n, k)$	Number of k-combinations of a given set with n elements
$\mathbf{crop}(\cdot)$	Image patch cropping function
$\mathbf{clothing}(\cdot)$	Clothing features extraction model
D	Diagonal matrix of node degrees from a graph
E	Set of edges from a graph
$\mathcal{E}$	Total number of edges in a graph
$\mathbf{emotion}(\cdot)$	Emotion features extraction model
$\mathbf{exp}(x)$	Calculates the value of e^x
F	Set of feature vectors
$\mathbf{F}(\cdot)$	Fourier transform
$\mathbf{g}$	Global node type from the SKG
G	Graph formed by a set V of vertices and a set E of edges
$\mathbf{gender}(\cdot)$	Gender features extraction model
$\mathcal{H}$	Hidden state size of the Social Graph Network model
I	Identity matrix
$\mathbf{I}$	Input image
$\mathbf{l}$	Local node type from the SKG
L	Normalized graph Laplacian matrix
$\mathcal{L}$	Total loss value

$\mathbf{LayerNorm}(\cdot)$	Layer normalization
$\mathbf{log}(x)$	Calculates the value of $log_e x$
$\mathbf{LSE}(\cdot)$	LogSumExp function
$\mathbf{M}(\cdot)$	Generic message passing function
$\mathbf{max}(\cdot)$	Calculates the max value of the inputs
$\mathbf{mean}(\cdot)$	Calculates the mean value of the inputs
N	Set of neighbors from a node
$\mathcal{N}$	Normalization factor for Non-Local Neural Networks
O	Set of object bounding boxes
$\mathcal{O}$	Total number of object classes for a given image
$\mathbf{object}(\cdot)$	Object features extraction model
$\mathbf{p}$	Personal node type from the SKG
$\mathcal{P}$	Total number of persons in a given image
Q	Set of queries
$\mathbf{r}$	Relation node type from the SKG
R	Set of social relationships
$\mathcal{R}$	Total number of social relationships in a given image
$\mathbf{R}(\cdot)$	Generic graph readout function
$\mathbf{ReLU}(\cdot)$	Rectified Linear Unit activation fuction
S	Set of social neighbors from a relationship node
$\mathbf{softmax}(\cdot)$	Softmax activation function
$\mathbf{sum}(\cdot)$	Calculate the sum of the inputs
$\mathcal{T}$	Total number of time steps for a message propagation method
$\mathbf{tanh}(\cdot)$	Hyperbolic tangent activation function
U	Matrix of eigenvectors ordered by eigenvalues
$\mathbf{U}(\cdot)$	Generic node update function
V	Set of vertices from a graph
$\mathcal{V}$	Total number of vertices in a graph
X	Set of input image patches

Contents

Chapter 1

Introduction

The increasing adoption of autonomous systems in several sectors of society raises the demand for technologies that will allow them to behave properly in such scenarios. For this reason, research topics associated with human analysis and behavior understanding have gained considerable attention recently. This type of knowledge can usually be employed for a variety of tasks, improving the performance of their applications.

In this context, a fundamental requirement to comprehend human comportment is being capable of identifying the main characteristics that define common social relationships. Although this is an essential topic for automated human analysis, it has not been adequately explored.

This work presents an approach to recognize social relationships between pairs of individuals using visual information from multiple scales, exploiting the interdependencies between relationships, and applying prior knowledge in the form of multi-scale attributes and other constraints to the model. The proposed method is able to achieve this by constructing a representation that preserves the structure of the social relationships from the input image and extracting the embedded information.

1.1 Contextualization

Recently, we have witnessed a rapid growth in the adoption of autonomous systems toward multiple aspects of our daily life, such as shopping, security, and transportation. For most cases, the devices involved in these tasks need to be capable of comprehending the environment around them, especially in situations where they are expected to coexist or even interact with human beings.

In these circumstances, it is essential to understand human behavior, since these systems might be exposed to conditions where their perception of human action can directly influence their performance, like on automated surveillance [Noceti and Odone, 2014]. Other situations, such as self-driving vehicles, can be even more critical, where

erroneous judgments may pose risks to nearby people [Brooks, 2017]. For these reasons, technologies capable of analyzing human behavior are fundamental, facilitating the coexistence between individuals and machines.

Considering visual data, a well-studied branch of the automated human-behavior analysis is the action recognition, where initial works mainly focused on individual actions [Herath et al., 2017] and later started to investigate pair [Gemeren et al., 2018] or group interactions [Wu et al., 2019].

Another frequently explored topic is emotion recognition, where we also have seen a similar shift from initial works focusing on single individuals [Ko, 2018], to group emotion [Guo et al., 2018]. The crowd analysis topic also emerged as a field of interest, with research on tasks such as person count [Li et al., 2015] and movement prediction [Yan et al., 2014].

Lastly, recent works proposed to extract even higher-level information that only makes sense within collectives, such as group identification [Varadarajan et al., 2017], group cohesion [Ghosh et al., 2018], group affect [Dhall et al., 2015] and social relationship recognition [Li et al., 2017], which is the topic of this work.

1.2 Motivation

As mentioned in the previous section, theoretically, every application that can be affected by human behavior is also able to benefit from social relation recognition. This section provides an explanation of how relation recognition can interact with other tasks, along with some examples of possible use cases in the future. The proposed applications take into account only the advantages provided by the consideration of social relationships, disregarding challenges related to image acquisition, pre-processing, and other specific technical issues related to these tasks.

Other works suggest that social relation recognition can be used to enhance the understanding of personal characteristics and also to help with behavior prediction [Smith and Zárate, 1990]. These aspects could improve the performance of a variety of computer vision tasks, such as group activity recognition [Ibrahim et al., 2015; Lan et al., 2012; Wu et al., 2019], image retrieval [Johnson et al., 2015], visual question answering [Teney et al., 2017], image captioning [Chen et al., 2020], scene graph generation [Raboh et al., 2020], social event recognition [Ramanathan et al., 2013], and family member identification [Dai et al., 2015].

The common ground between all the mentioned tasks is that they benefit from better image understanding, and when there are persons involved, recognizing social re-

lationships can be a key feature, since humans tend to interpret a scene based on social interactions between the individuals depicted [Guo et al., 2019]. For example, relation recognition can enhance image captioning by generating more precise descriptions with the inclusion of words describing specific person or relation traits. In this sense, the image from Figure 1.1a could be captioned more precisely as "Family picnic" instead of a generic description such as "Group of people having a picnic".

Another example is Human-Robot Interaction (HRI), where behavior analysis becomes imperative to natural interactions between humans and machines [Kong and Fu, 2018]. For this purpose, social relation recognition can be employed in an attempt to read the environment around the robot, allowing it to react in a socially adequated manner [Bartlett et al., 2019]. The social robot Pepper depicted in Figure 1.1b implements a similar concept employing emotion recognition to select a suitable approach to initiate an interaction.

(a) Family picnic

(b) Social robot

Figure 1.1. (a) An image captioning sample that can be more precisely described by considering social relationships. (b) Pepper is an example of a social robot that can employ emotion recognition to interact adequately [Alecrim, 2015].

A possible application of these social robots is for customer assistance in a commercial establishment. This is the case for Gal (Figure 1.2a), a robot created by Gol airline that was used to assist passengers at Guarulhos Airport [Casagrande, 2019]. She helped by answering questions and guiding persons between several areas of the airport while also providing entertainment. Another company that adopted a similar strategy was Bradesco. The bank employed a robot named Link237 (Figure 1.2b) on customer service for an agency in São Paulo.

(a) Gal **(b)** Link237

Figure 1.2. (a) Gal was presented by Gol at Guarulhos airport to assist passengers [Casagrande, 2019]. (b) Link237, the customer service robot introduced by Bradesco for a bank agency in São Paulo [Daraya, 2013].

1.3 Problem

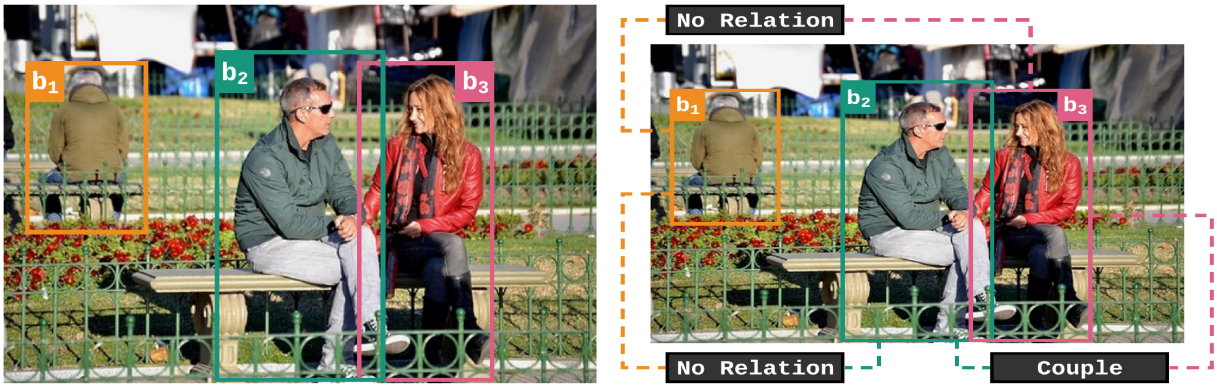
Social relationships can be defined as the connections between people who have recurring interactions, which are perceived by the participants to have personal meaning [August and Rook, 2013]. Early methods were capable of inferring social relationships from different information sources such as text [Fairclough et al., 2003] and images [Singla et al., 2008]. This work is focused on social relation recognition from visual data, which can be defined as the process of identifying the types of connections between each pair of individuals on a given image.

The relation classes considered are not directly tied to the problem. Instead, they are dependent on the social theory employed to understand the relationships on each database. The task consists only of reproducing the human perception of these relations, which is reflected in the labels provided by the benchmarks. Additionally, social relation recognition is an ill-posed problem, since the information necessary to correctly identify some relationship classes may not be present in the image, subjecting them to multiple interpretations.

Since the task manipulates visual data, it is also affected by most of the typical computer vision issues, such as variations in scale, appearance, illumination, and pose [Li et al., 2017], which can heavily impact the performance of the system. These effects are kept out of the scope of this work, and for this reason, the images from the databases employed are usually collected from social media, avoiding distorted pictures, crowds, and any other type of noise that could interfere with the results.

Additionally, the benchmarks employed in the experiments also provided the bounding boxes for the considered individuals, dismissing the need for person detection

and allowing this work to focus on solving challenges specific to the social relation recognition task, which are described in this section. An input image example depicting three individuals is shown in Figure 1.3, along with the corresponding relationship classification for each pair.



(a) Input **(b)** Output

Figure 1.3. An example of the social relation recognition task. (a) Input image provided along with the corresponding bounding boxes for each person. (b) The relationship between each pair is classified.

One of the most recurrent problems in supervised relationship recognition is that the models trained with raw data can sometimes learn wrong associations [Sun et al., 2017], behaving as expected with the training data but hardly generalizing to other situations. In this work, the hypothesis that this happens because they do not actually reflect how we, as humans, perceive social relationships is evaluated, which is an essential characteristic when performing this task, since the social relation concept itself is rooted in human perception.

The way how humans differentiate between each relationship type is profoundly related to appearance attributes such as age, gender, clothing, emotion, and body positioning, as proposed by research psychologists [Bugental, 2000]. For example, intimate relations tend to show a closer body distance, and if combined with significant age differences, it can indicate a parent-child relation. The influence of these attributes was explored by initial works on social relation recognition, but alone they are not enough to correctly identify relationships, which are also dependent on higher-level information (e.g., action, emotion, context) [Li et al., 2017].

Context data can also play an essential role in defining social relationships, helping to reduce the ambiguity present in situations where similar arrangements may represent distinct types of relationships. For example, the presence of office or household items can differentiate between professional or friendship relations [Wang et al., 2018b], as shown in Figure 1.4.

Other relationships on the same image are also a meaningful source of information, since they follow strong logical constraints, which can be exploited to recognize social

(a) Business reunion **(b)** Friends meeting

Figure 1.4. Similar arrangements can be differentiated with the help of context. (a) Business reunion identified by office objects. (b) Friends meeting denoted by household items.

relations [Li et al., 2020], as shown in Figure 1.5a. In this case, if we know that the relationships in blue are *parent-child*, it becomes easier to infer that the third relationship is *couple*.

Finally, since social relationships are defined pairwise, it means that the total number of relationships $\mathcal{R}$ for an image containing $\mathcal{P}$ persons can be calculated by

$$\mathcal{R} = \mathbf{C}(\mathcal{P}, 2) = \frac{\mathcal{P}!}{2!(\mathcal{P}-2)!}, \tag{1.1}$$

and for images with significant numbers of persons, it scales to a really complex problem, not only because of the total number of relationships, but also because we have to take into consideration how they affect each other, and most importantly, how the classification can change for the same individual depending on the person we are pairing with, as shown in Figure 1.5b.

(a) Small family **(b)** Multiple relationships

Figure 1.5. Samples containing dependencies between relationships and multi-relation arrangements, where the numbers identify each individual and the arrows represent their relations. (a) *Couple* relationship in red and *parent-child* relation in purple. Any relationship in this example can be inferred from the other two. (b) *Couple* relationship in red, *friends* relationships in green and *commercial* relationships in blue. Persons 1 and 2 are part of three different relationship types in this example.

Previous works try to overcome the mentioned challenges by exploiting some proprieties of social relationships. However, they focus on specific aspects of the problem, such as semantic attributes or environment objects, often overlooking the interaction of multi-scale information and the role of other social relationships in the recognition process. More specifically, they do not consider the structure of social relationships as a whole but instead only smaller portions of the problem independently.

In this work, the concept of social structure is investigated, considering the hypothesis that all mentioned factors play a crucial role in the recognition process by offering complementary information that must be fully acknowledged to achieve an adequate understanding of social relationships. The conducted analysis aims to provide the foundations to develop a framework capable of fully representing all the relevant information, hence preserving the original structure of these relations. Finally, this representation can be exploited for the social relation recognition task, bringing the reasoning process closer to how humans perceive social relationships, which increases the performance over previous techniques.

1.4 Contributions

The main contributions presented in this work are threefold, starting with the proposition of a novel approach to interpreting social relation recognition methods, which is based on the scope of the employed information, the ability to apply prior knowledge in the form of model constraints, and the consideration of interdependencies between different relationships in the same image. The proposed taxonomy provides a framework that encompasses all previous works, being capable of distinguishing the main strengths and shortcomings of each method.

A new representation for social relationships is introduced to cover the deficiencies identified in previous works, the Social Knowledge Graph (SKG). This structure is able to carry all the pertinent information for social relationships while applying prior knowledge and other meaningful constraints to the model. More specifically, it can represent learned features, pre-trained attributes, and how this information is combined for multiple scales, achieving a complete portrayal of social relationships. The dependencies between all these types of information form a structure, which is broken by previous works when they separate relationships pairwise or when they consider information from a particular scale alone. However, this structure is replicated in the proposed graph representation, offering the reasoning model all the information required to achieve the high-level understanding necessary for the social relation recognition task.

The final contribution is the Social Graph Network (SGN), a deep graph model that implements three distinct message propagation methods to exploit the unique and intuitive graph structure. Each proposed spatial convolution operation is specifically designed to extract information from a particular region of the graph, optimizing the reasoning process and selecting the most meaningful features via knowledge propagation. Additionally, the approach also employs mechanisms on representation and reasoning levels to reduce the noise generated by the interaction between all this data. The final result is a high-level description for each social relationship containing multi-scale and multi-attribute information, which also incorporates dependencies from other relationships, according to the graph structure.

As revealed by the proposed taxonomy, no previous work was able to combine information from all scales simultaneously, nor to add attributes or apply other constraints to multi-scale data. They are also unable to capture correlations between relationships, treating them as independent events, which breaks the social structure presented in the image. The proposed approach incorporates all these proprieties, improving the model performance by bringing the process closer to how humans perceive social relationships.

1.5 Work Organization

The remainder of this work is organized as follows: Chapter 2 introduces a new approach to interpret social relation methods, and Chapter 3 presents a literature review including all available relationship recognition works employing image and video data. The main concepts and methods related to deep graph neural networks are explained in Chapter 4, while the proposed approach is detailed in Chapter 5. Finally, the conducted experiments along with their results are discussed in Chapter 6, and Chapter 7 concludes the work, followed by future research directions in Chapter 8.

Chapter 2

Taxonomy

In order to contextualize the contributions of the proposed approach, a taxonomy for social relation tasks is presented in this chapter, aggregating practical knowledge from previous works, and also incorporating social theories introduced by multiple authors. The criteria applied in the categorization process are based on three main concepts: information scales, data dependencies, and model constraints.

These conceptions are derived from fundamental aspects of the social relation recognition problem mentioned in the previous chapter, and they are further elaborated in the following sections. Together they constitute a framework that not only allows the comparison between all previous social relation recognition methods, but also provides means of identifying their shortcomings. The insights obtained from this analysis were the driving factors behind the design choices made during the elaboration of the method proposed in this work, which is formally described in Chapter 5.

2.1 Social Scales

Here, a hypothesis describing how information from different perspectives influences the interpretation of social relationships is introduced. This proposition motivates the adoption of three distinct social scales, namely, personal, local, and global scales. Most previous works are able to exploit, in some form, data obtained from only one or two of them. Additionally, they also lack clear explanations on why these scales work for social relation recognition and the design choices behind their application.

More specifically, the main benefits for using each of these information scales are presented in this section, along with an analysis of their descriptive power, demonstrating how they offer not only meaningful but also complementary information. For this reason, all of them need to be combined to achieve information completeness for social relation recognition tasks, which is done by the method proposed in this work.

A common technique employed in computer vision problems is the extraction of

information using different scales, where they can assume distinct meanings depending on the context [Lindeberg, 1994]. The scale-space theory is a very successful framework capable of exploring this concept using handcrafted techniques. For deep learning, a possible approach that abstracts the idea of different scales is implemented by cropping patches from specific regions of the image, reducing the noise by removing unwanted information and limiting it to a target context, from which the model can learn.

Methodologies with analogous concepts became the state-of-the-art for human-behavior problems, such as group activity [Ibrahim et al., 2015] and group emotion recognition [Guo et al., 2018]. For social relation recognition, the situation was no different, as many previous works [Li et al., 2017; Wang et al., 2018b; Yan and Song, 2019; Zhang et al., 2019] combined local and global information with the purpose of incorporating context to a specific relationship, which can help to distinguish between some ambiguous situations, as mentioned in Section 1.3. To explain why these approaches work for social relation recognition, it is necessary to understand how the information from these distinct scales is correlated from a social point of view.

The most fundamental components of our societies are the individuals, and from their interactions, groups and communities emerge, giving form to social relationships [Barkan, 2011]. In this sense, it is reasonable to imagine that the characteristics of a group of people are a product of their individual traits in a bottom-up manner, and it is also possible that individual actions are influenced by the behavior of the group as a whole, in a top-down direction [Barkan, 2011].

By exploiting this social structure and the relation between its parts, works that deal with human behavior are able to approach the same problem from multiple perspectives, defining different objects of analysis (e.g., individuals, pairs, groups, crowds). Although social relationships are interpersonal, they are also heavily influenced by this social structure, where individual and group aspects play an essential role, as explained in Section 1.3. For this reason, all the regions of the image corresponding to those subjects also need to be considered.

Previous works extract features from multiple arbitrary areas, including face [Zhang et al., 2015; Guo et al., 2019; Sun et al., 2017; Wang et al., 2020], body [Li et al., 2017; Goel et al., 2019; Li et al., 2020], pairwise [Wang et al., 2018b; Goel et al., 2019], and even the full image [Zhang et al., 2019; Liu et al., 2019]. However, different types and volumes of information are captured from each one of these image regions, which can heavily impact the outcomes, as illustrated in Figure 2.1.

Motivated by the described scenario, this work introduces a methodology to interpret social relation problems based on the image regions extracted from three different scopes named *social scales*. Each one encapsulates a specific type of information associated with a distinct perspective from social relationships, providing unique benefits and disadvantages, while offering complementary data. The proposed social scales can be

Figure 2.1. An example of how changes in information volume and detail provided by different image regions can influence the outcome of a classification. Face images offer detailed but insufficient information, which is usually applied to infer important individual features such as age and gender, which could lead to a *couple* classification. Pairwise body-focused pictures provide individual and relative features such as clothing, pose, proximity and interaction, probably changing the classification to a *coworkers* relation. Finally, in this case, only the full picture was able to present critical background information, such as the divan, to infer the correct *commercial* relation between a psychologist and her client.

described as:

Personal This scale carries mostly personal information, contained in body region image patches, obtained by cropping every person separately. The main objective is to restrict information to an *individual* scope, removing all the noise generated by the rest of the image while preserving fine-grained appearance details and allowing the extraction of person-specific attributes such as age, gender, clothing, and pose. Naturally, the scope restriction severely impacts the information obtained from the background, as shown in Figure 2.2, incapacitating the model to learn other types of features that can be important to determinate social relationships, requiring the combination with other more far-reaching social scales.

Figure 2.2. Personal-scale image patches restrict the information to a *individual* scope and allow the extraction of relevant personal features. However, it is impossible to infer proximity and activity from these image regions while the background and surrounding objects are completely ignored, dismissing these important sources of information.

Local The local scale reflects pairwise *relative* information, restricting the image to the smallest region containing both individual scopes for each pair of persons taking part in a relationship, providing an uninterrupted point of view connecting these

persons, which allows the extraction of crucial features such as activity and proximity. Although this model provides insight into background information, it is not enough to present a full context, possibly overlooking essential data. This approach also has a significant drawback when other persons are in between the current pair, making these individuals appear in the cropped image patch, as shown in Figure 2.3. In this work, these occurrences are referred to as *interferences*, and when they negatively affect the outcomes, undermining the scope restriction concept and generating noise during the feature extraction process, they are called *destructive interferences*. Another related problem happens for a group of persons who are too close, resulting in many of the obtained image regions being similar to each other, even for distinct pairs, which can have different labels. These occurrences result in inconsistencies which can possibly impair the learning process if personal-scale information is not provided.

Figure 2.3. Local-scale image patches are adequate to extract pairwise *relative* features, allowing the estimation of distance and interaction between two individuals while also providing insight into their background. However, it is possible to see from Local 1-3 that this scale has a significant disadvantage when there are persons in between the considered pair, resulting in *interference*, which undermines the scope restriction. A full context is also not provided by this scale, which may result in neglecting important environmental information.

Global This scale takes into account *general* information, which can affect all individuals and relationships in the same image. The scope is defined by the whole image, as shown in Figure 2.4, and it provides data that can be decisive to obtain a correct classification, such as the environment or even objects that are too distant from individuals to be captured by other scales. It allows the model to learn relationship-agnostic features that can be combined with data from other scales, providing additional contextual information. The evident disadvantage of this scope is the lack of specification on the individuals participating in the current relationship, considering the same image for every pair in the picture. However, in some cases, the model will be provided with different labels for these relations, hindering the learning process if this data is not associated with other individual-specific information.

Likewise typical local and global features employed in other computer vision tasks, each social scale has its benefits and shortcomings. The closer the scale, the more specific

Figure 2.4. Global-scale images provide *general* information, which can influence all the relationships contained in the image. It can help to identify critical objects and environment cues out of reach from other scales. However, it is relationship-agnostic, generating features that are inadequate for identifying relationships alone, serving only as a complementary information source to other scales.

the extracted features, and the less they are affected by noise [Hassaballah et al., 2016]. However, these features also carry less context information, which may be essential to the problem in some cases. It is also important to reinforce that the social scale concept introduced in this section refers to the scope of the information and has no direct relation with the size of the image regions.

However, another factor to consider for deep-learning-based methods, or any other technique that downscales the input images, is that they can exacerbate the proprieties related to each social scale. This effect occurs because the lower scales are just smaller portions of bigger scales image regions, which means they do not have to be reduced as much to reach the adequate input size, preserving more information.

For example, when personal and local-scale image patches are downscaled to the same dimensions, the capability of the local scale to preserve fine-grained features is diminished even more, while personal-scale patches are able to retain some details. This means that social relation recognition methods that employ scale variant approaches lose even more information when using a smaller subset of the proposed social scales.

All the information necessary for social relation recognition tasks can be obtained by employing these three scales, generating a complete set of features covering the subjects involved in relationships from all relevant social perspectives. More specifically, the personal scale encompasses *individual* data, the local scale covers *relative* information between each pair of persons, and the global scale comprises *general* data.

2.2 Data Dependencies

This section explains the data dependencies between relationships, which constitutes a significant source of information that can be exploited to increase the performance of the model. Some examples are presented to illustrate this concept while also offering insight into how it is implemented in this work.

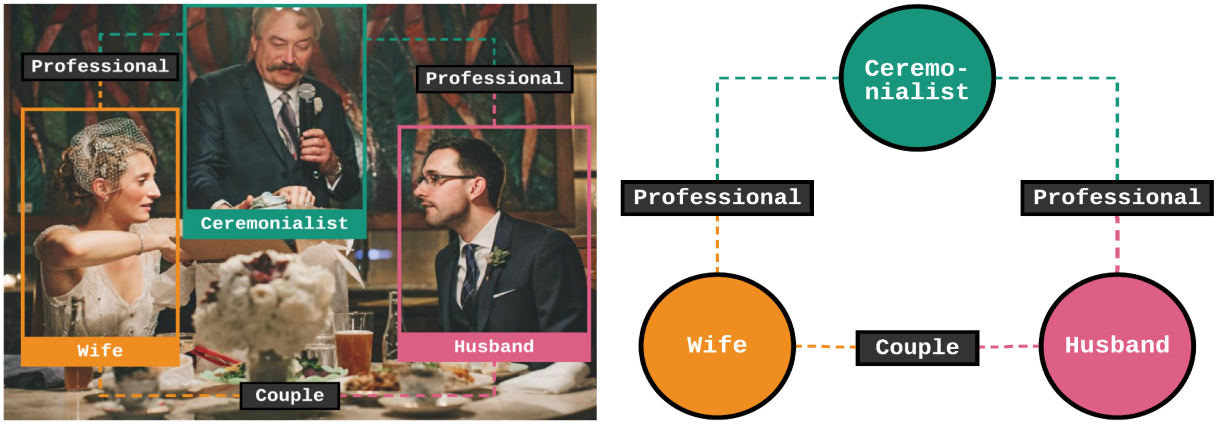
Previous social relation recognition methods focused their efforts on experimenting with different combinations of information scales and prior knowledge, mainly in the form of appearance attributes. All of these works consider only features from the individuals taking part in the relationship, classifying each pair separately from the others in the image. This kind of approach handles social relationships as independent events, but in some cases, it is possible to infer a relation based on information from other pairs. This means that social relationships are instead correlated, and this propriety can be exploited to improve the model.

More precisely, previous methods seek data dependencies only between the information extracted from the pair of individuals participating in the current relationship. In this work, these are considered as *intradependencies* since they are intern to the social relation. However, relationships are also correlated with each other, and these types of data dependencies are referred to as *interdependencies*.

A knowledge graph is a suitable option to represent not only the concept behind these interdependencies, but also the entire structure of the relationships depicted in the image, which allows the model to access information from other pairs, helping it to learn their correlations. An example image is shown in Figure 2.5, along with the corresponding knowledge graph, representing each person with a node and the social relationships between them as the labeled connections. It is important to mention that this is not the final version of the graph proposed in this work, and the only objective of this image is to illustrate the concepts introduced in this section.

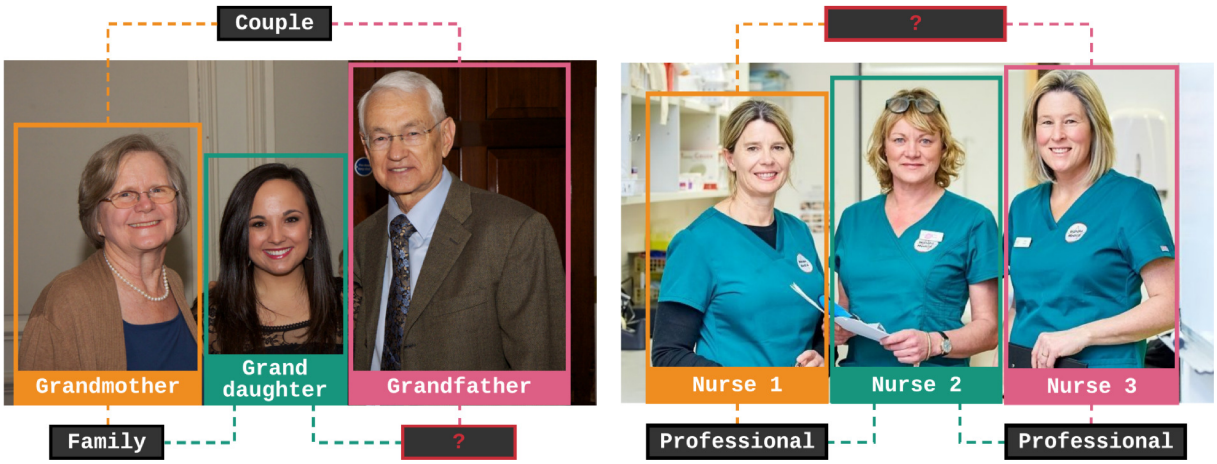
Considering the structure depicted in Figure 2.5b, it is possible to use the labels of known relations with other individuals to predict an unknown relationship. This is exemplified in Figure 2.6b, where the *couple* relationship between the two senior persons is given, and it is also known that the young girl is the granddaughter of the lady. In this scenario, it is trivial to infer that the man also has a *family* relationship with the girl; more precisely, he is her grandfather. Another example is shown in Figure 2.6a, but in this case using *professional* relationships, where it is also possible to predict the relation between the pair of women as work colleagues since they both have the same relation type with a third person.

However, to exploit these interdependencies, it is necessary to handle the image as a whole, and this is not done by previous works. They consider relationships independently,



(a) Example image **(b)** Knowledge graph

Figure 2.5. A representation showing the concept of how relationships form a structure within an image. The knowledge graph is capable of preserving this structure, which can be employed to infer unknown relations using information from other known relationships.



(a) Family relation **(b)** Professional relation

Figure 2.6. Examples of the inference of unknown relationships, using *interdependencies*. (a) In this scenario, if the senior lady is the grandmother of the young girl and also the wife of the senior man, then it is possible to conclude that the man also has a *family* relationship with the girl, being her grandfather. (b) A trivial situation where the two women in question have the same *professional* relationship with a third woman, meaning they are also coworkers.

examining only the features for each pair of individuals, which breaks the social structure depicted in the image. In this work, these dependencies are captured by the proposed knowledge graph, replicating the structure of the social relationships, which can be used to perform the presented inference process. In addition, attributes interdependencies from multiple scales are also captured by the model, allowing the acquisition of even higher-level correlation between relationships.

2.3 Model Constraints

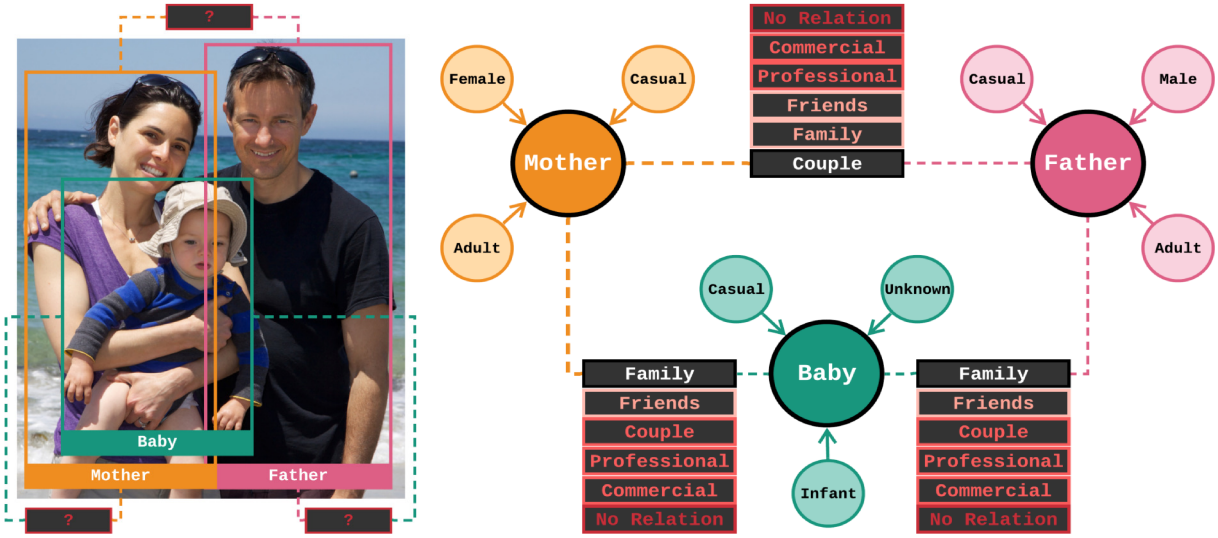
In this section, the dependencies between appearance attributes and social relationships are investigated. It is shown that these correlations can sometimes be immediately inferred, but in other cases, they are tied to cultural aspects. Next, it is presented how these data dependencies work for social relation recognition and how they can be exploited to enhance the model, along with other types of prior knowledge, resulting in associations that are more similar to how humans identify social relationships.

As mentioned previously, research suggests that appearance attributes such as age, gender, clothing, emotion, and body positioning play an essential role in how humans perceive social relationships [Bugental, 2000]. This means that if we want to build systems capable of recognizing social relationships, they need to make similar associations since the concepts they have to learn are rooted in human interpretation.

Sometimes, the dependencies between social relationships and these attributes can be derived directly using simple deduction. For example, in parents and children *family* relationships, the parents have to be older than their children. It does not matter what kind of *family* is being considered, the parents will always have a higher value for the age attribute. In other cases, the relationship can have a strong correlation with some attributes, or they may even be tied to culture. Some examples would be *couple* relationships, which tend to happen between people of opposite genders, and the use of formal clothing for *commercial* and *professional* relations, which is a cultural trace. Either way, even if these dependencies are not hard constraints, as in the former age example, all of them can be exploited to improve the model, as illustrated in Figure 2.7.

Most of the previous works add prior knowledge to their models by considering multiple combinations of attributes, since they cannot be learned directly from the training data. The available benchmarks only provide relationship annotations, which means that pre-trained models have to be employed to obtain this type of information. For this reason, attribute feature vectors are fixed, acting somewhat as inputs to the model and serving as constraints to the domain by defining the traits that have to be considered for the task. The model works on strategies to aggregate these attributes while also learning new features associated with them directly from the image. This effect is similar to well-known task interactions for multi-task models, where the joint-learning of multiple related tasks can help to increase the model performance, and for some applications, the results can improve continuously with the number of tasks [Ruder, 2017].

This information interaction is a powerful tool that can be implemented using the knowledge provided by previous research on multiple fields, ranging from studies explaining how humans interpret each type of attribute [Bugental, 2000] to works analyzing the role certain traits play in social relation recognition [Sun et al., 2017]. By adding prior



(a) Typical family photo **(b)** Knowledge graph with individual attributes

Figure 2.7. An example of attributes acting as constraints to the classification. (a) Image with three unknown relationships. (b) The corresponding knowledge graph considering attribute nodes for age, gender, and clothing. The probability of each relationship class depends on the interactions of these attributes values for each pair. Age and gender increase the odds of a *couple* relationship between the mother and the father, while casual clothing reduces the chances of *professional* and *commercial* classes. The age factor is also crucial to identify the *family* relationship between the parents and their baby.

knowledge, the decision-making process can get closer to the human perception of relationships, which is a desirable aspect for these models, since social relations are human constructions interpreted from a human perspective. Additionally, this measure can help to solve the generalization issues reported in Section 1.3.

In this work, attributes and any other type of prior knowledge that can be associated with a specific scope tied to a social scale are considered as being model constraints. In other words, they are pre-defined information arbitrarily enforced to the model in an attempt to guide the learning process in a particular direction by capturing the dependencies between learned and given data. For example, extracting age features, using a pre-trained model, from individual body image patches, which are linked to the personal scale, or extracting group activity features from local-scale pairwise image patches. Object features from a scene are also considered as attributes, since they contain specific pre-learned information extracted from smaller portions of the image. This definition allows a better categorization of previous works that employ attributes and other types of prior knowledge or constraints to their models.

Chapter 3

Related Work

In this chapter, an overview including the main contributions of other works on social relation recognition is presented, along with an analysis on how the approach proposed in this work differs from these methods, specifying strengths, weaknesses, and other characteristics. For this purpose, the concepts presented in the previous chapter are applied, allowing to draw a parallel between these techniques by considering fundamental aspects of social relationships.

Most works in Computer Vision concerning human social behavior usually deal with two different problems. The first one is called Social Relation Recognition, and it consists of identifying the type of a relationship. This is a typical multi-class classification task, where the considered relationships vary according to the methodology selected to build the benchmarks, as stated in Section 1.3. The second is a binary classification problem denominated Social Relation Trait, and it involves detecting multiple traits from a relationship, such as dominant, competitive, trusting, warm, friendly, attached, demonstrative and assured. Although these are different problems, their similarities facilitate the exchange of meaningful insights, justifying their inclusion in this chapter.

3.1 Social Relation Recognition

The social relation recognition problem has a recent formulation, and for this reason, there are few works exploring this task. Here, the taxonomy defined in Chapter 2 is applied to contextualize all of the previous methods on social relation recognition present in the literature. They are defined in terms of how the concepts of social scales, relationships interdependencies, prior knowledge, and model constraints are employed.

Previous works also established that social relationships tasks can be performed in image and video data. Although this work focuses on still images, a review of video approaches is also presented, considering they can provide important concepts and ideas.

3.1.1 Images

Li et al. [2017] presented a benchmark for social relation recognition inspired by the relational theory [Fiske, 1992], defining hierarchical relationships from coarse (3 classes) to fine (6 classes) levels, with the name of People in Social Context (PISC) dataset. It considers not only family relations, but also a set of other relationship classes that are claimed to cover all aspects of human interactions.

For the recognition task, they proposed a convolutional model employing personal and local-scale information, obtained by extracting features from body region patches separately, and fused in pairwise images. The input patches are combined with bounding box coordinates, resulting in a single feature vector for each pair containing multi-scale information. Global information is addressed in the form of object attributes, which are detected using a Faster R-CNN model [Ren et al., 2015]. The features extracted from the object regions are weighted, applying an attention mechanism, and then fused with the vector previously obtained, generating the final set of features used for the classification, as shown in Figure 3.1.

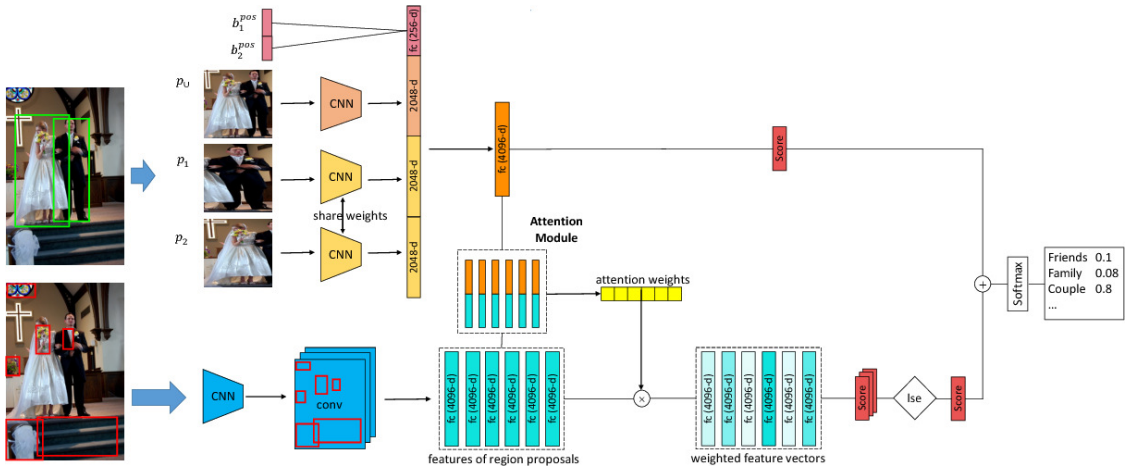


Figure 3.1. An overview of the Dual Glance model [Li et al., 2017], which is able to learn individual and relative information, also adding global context as object attributes, weighted by an attention module.

This work was essential for social relation recognition, not only because of the benchmark provided, which is one of the most used recently, but also due to the concept of adding context with object features, and the approach aggregating individual and relative information. However, this method is unable to learn global features directly from the input or to set constraints on other scales, and it also neglects relationships interdependencies, considering each pair separately.

Sun et al. [2017] was another relevant social relation recognition work, also presenting a new dataset named People in Photo Album (PIPA) relation, backed by the

domain-based theory [Bugental, 2000], which divides social life into 5 domains that are used to derive 16 relationship classes. It proposed a method that only made use of personal-scale information, employing individual face and body image patches to extract 12 semantic attributes while also measuring their contribution to the final performance. The obtained results correlate with the predictions of the domain-based theory, as presented in Figure 3.2.

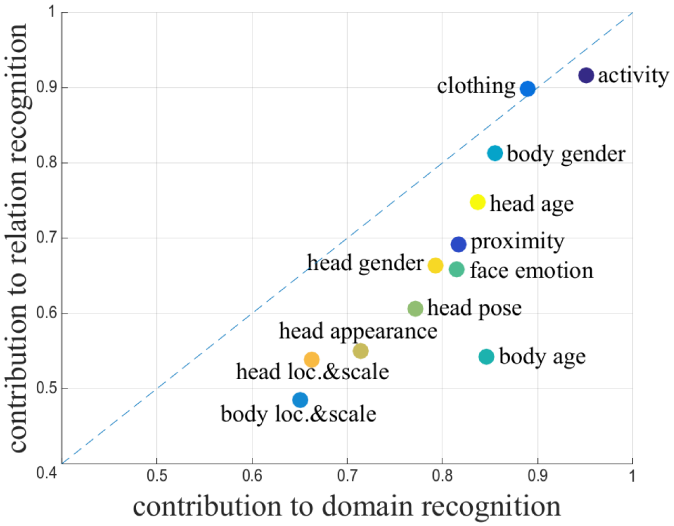


Figure 3.2. The contribution of 12 semantic attributes for the classification accuracy on the domain (5 classes) and relation (16 classes) splits of the PIPA-relation [Sun et al., 2017] dataset.

Their proposed model consists of pre-trained double-stream convolutional neural networks (Figure 3.3) to extract pairwise features for each attribute, which are combined in various forms and fed to an SVM classifier. Although their work provides significant contributions by making the benchmark and attribute models available, it does not bring new relevant technical aspects.

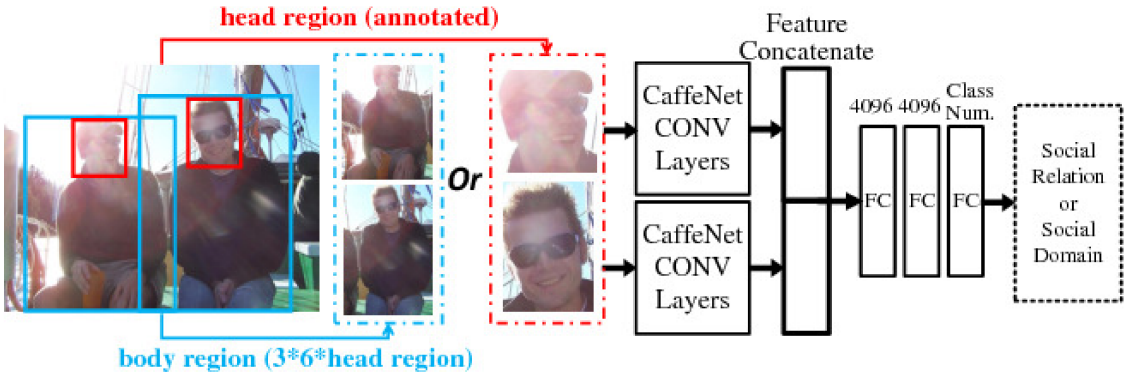


Figure 3.3. The double-stream CaffeNet model is composed of two convolutional networks sharing weights to learn specific pairwise attribute features.

Wang et al. [2018b] proposed a model that fuses individual and relative information with global-scale attributes while exploring prior knowledge indicating co-occurrences between objects and relationships classes, which are represented by a graph structure

for the whole dataset. The features are extracted in the same way as Li et al. [2017], but instead employing a pre-trained VGG 16 model [Simonyan and Zisserman, 2014] to classify the detected object regions. The main difference in this work is how the multi-scale feature vectors are combined with the object attributes, which is by employing a Gated Graph Neural Network (GGNN) model [Li et al., 2016] to propagate them through the graph structure, also applying an attention mechanism to objects classes, generating a final set of aggregated features used for the classification, as shown in Figure 3.4.

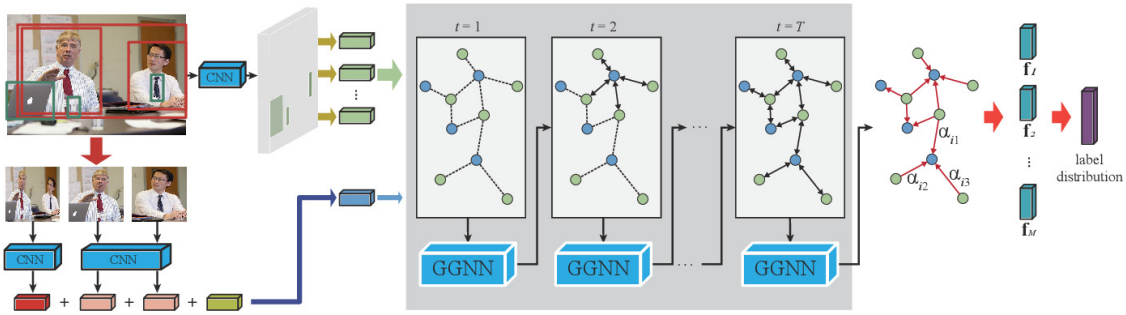


Figure 3.4. The Graph Reasoning Model (GRM) [Wang et al., 2018b] is very similar to the Dual Glance [Li et al., 2017], even employing the same methods to extract individual and relative features. The main difference is how the multi-scale feature vector is combined with global object attributes, which is done using a Gated Graph Neural Network (GGNN) [Li et al., 2016] guided by a graph structure representing class co-occurrences for the whole dataset.

The proposed Graph Reasoning Model (GRM) is very similar to the Dual-Glance model [Li et al., 2017], and for this reason, it also presents the same issues. Furthermore, the only difference lies in the method for combining the extracted features, which adds complexity to the model for a slight improvement since the only function of the graph structure is to represent class co-occurrences. The way how this information is obtained also restricts the approach since these values are pre-calculated and have an arbitrary threshold, which could be instead learned from the data.

Similar to the previous approach, Zhang et al. [2019] made use of graphs, but instead to represent person poses and their interactions with surrounding objects. This information is obtained from personal and local-scale regions for every pair of persons, employing CNNs for feature extraction and simple baseline models [Xiao et al., 2018] for pose estimation, while objects are detected using a Mask R-CNN [He et al., 2020]. It also extracts global-scale features from the whole image using a ResNet 101 [He et al., 2016] model, while two GCNs [Kipf and Welling, 2017] are employed to aggregate person-object interactions and pose data, represented by two separated graph structures. The outputs from the GCN layers are used to generate class scores, which are fused to global features scores, producing the final predictions, as shown in Figure 3.5.

The person-object graphs are an interesting contribution, providing information about the interaction between persons and objects. However, the late fusion design is a problematic approach when classifying global-scale images alone, as mentioned in Sec-

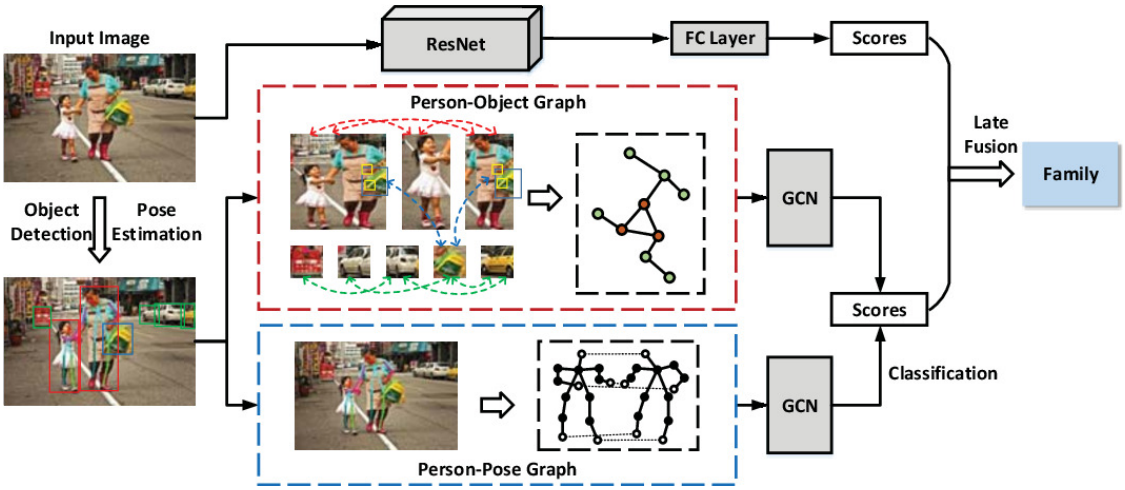


Figure 3.5. The Multi-Granularity Reasoning (MGR) model [Zhang et al., 2019] employs prior knowledge in the form of two graph structures: person-object graphs and pose graphs, which are processed using two GCNs [Kipf and Welling, 2017]. It also incorporates global features extracted from the whole image using a ResNet 101 [He et al., 2016] model. The class scores obtained from graph and global information are combined using late fusion.

tion 2.1. In this case, the network will receive distinct labels for the same image, generating inconsistencies, which apparently have been compensated by the scores obtained from the graph networks.

Goel et al. [2019] uses only personal (e.g., age, gender, clothing) and local (e.g., context and activity) scale attributes obtained from body and pairwise image patches. These features are extracted employing the pre-trained models provided by Sun et al. [2017] and fed to GRU cells [Cho et al., 2014], which aggregates them into a final representation. The model applies a multi-task loss to exploit a small set of age and gender annotations published by Oh et al. [2020], as shown in Figure 3.6. The main contribution offered by this work is the concept of learning social relationships and attributes jointly to generate a social relation graph based on image-graph generation tasks.

However, in this case, the graphs are not used as tools to represent social relationships for the reasoning process and instead, they are the product of the model's predictions. Another issue is the very restricted number of age and gender attribute labels for the PIPA-relation dataset [Sun et al., 2017], which are not enough to generalize to other datasets. Finally, the model is also unable to learn features directly from the input images.

Wang et al. [2020] proposed a feature selection method for social relation recognition, which measures the contribution of personal attributes extracted from face and body image regions. The model considers gender, appearance, emotion, pose, scale, clothing, activity, and proximity attributes obtained using the pre-trained networks provided by Sun et al. [2017]. The proposed attention mechanism learns weights for specific types of attributes, providing insight on how each one impacts the performance of the model, which is illustrated in Figure 3.7.

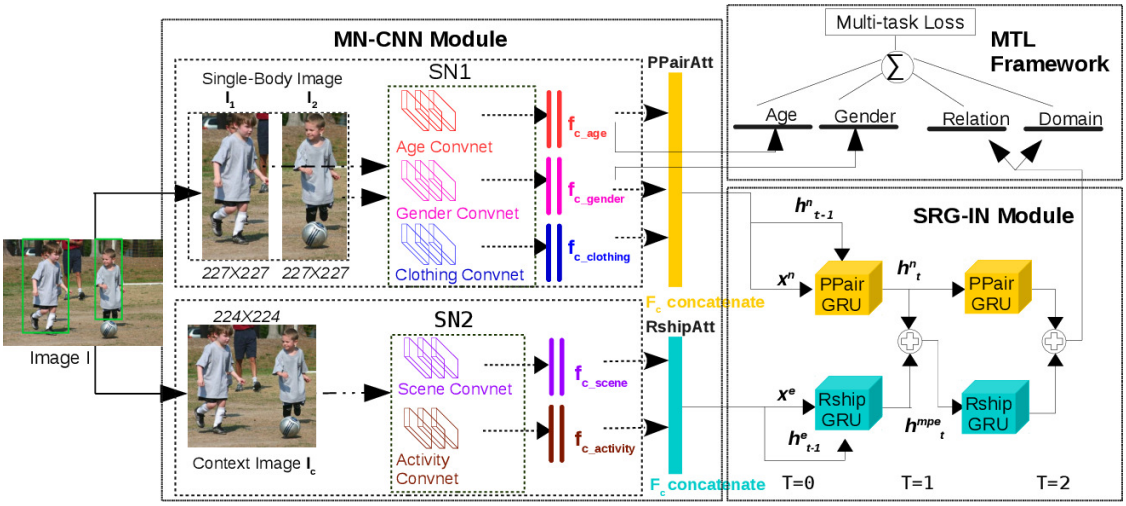


Figure 3.6. The Social Relationship Graph Inference Network (SRG-IN) [Goel et al., 2019] employs age, gender, clothing, activity, and context attributes extracted with pre-trained models and combined using pairs of GRU cells [Cho et al., 2014].

Figure 3.7. The Deep Supervised Feature Selection (DSFS) method proposed by Wang et al. [2020] to measure the contribution of multiple attributes extracted from face and body image regions, employing the models provided by Sun et al. [2017].

The purpose of this work is to extend the analysis done by Sun et al. [2017], further investigating the contribution of each attribute, and also providing a mechanism to select the most meaningful ones. For this reason, it suffers from the same problems as its predecessor, offering only small technical contributions besides the feature analysis.

Finally, Li et al. [2020] is the state-of-the-art for social relation recognition and the only image-based method capable of capturing relationship interdependencies. This is done by employing a ResNet model [He et al., 2016] to learn features from body image patches, which are extracted applying ROI pooling [Girshick, 2015], and used to build a graph representing each person from the input image, as shown in Figure 3.8. The structure is fed to a GGNN [Li et al., 2016] model, which aggregates the extracted features that are finally used to predict the existence of relationship edges between each pair of nodes.

This model is the first to consider information from other relationships in the

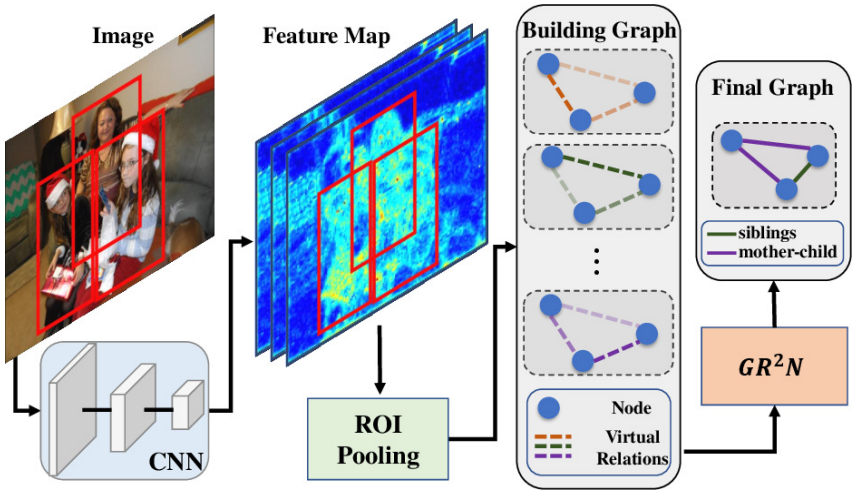


Figure 3.8. The Graph Relational Reasoning Network (GR^{2}N) [Li et al., 2020] generates a graph where each person is represented by a node and predicts the existence of a relationship edge between them. This is done by aggregating data from each pair with a GGNN [Li et al., 2016] model. These features are obtained by applying ROI pooling [Girshick, 2015] in the feature maps of the extraction backbone.

same image, but this is done in a very restricted manner. The method is only capable of dealing with personal-scale information, completely neglecting relative information from the local scale, and especially the global scale, which makes this work blind to context. In addition, the design choice to treat social relation recognition as an edge prediction is somewhat unnecessary and could be simplified to a edge classification problem since the graph is always complete. These shortcomings are reflected in the model performance, which overcomes previous works using pairwise approaches by only a small margin.

3.1.2 Videos

The following methods focus on social relation recognition from video, which is a different problem from the one considered in this work. For this reason, only a brief description including the main concepts behind each work is provided, with the primary purpose of showcasing methods and concepts that could also be implemented for image-based approaches.

Lv et al. [2018] extracts only global-scale information, using the entire frames from sampled video segments. The proposed model also manipulates multi-spectral data, including spatial-temporal features from the video frames and optical flow, extracted with the help of a TSN model [Wang et al., 2016b], and audio spectrum features obtained with a GoogleNet [Szegedy et al., 2015]. All this information is combined employing a late

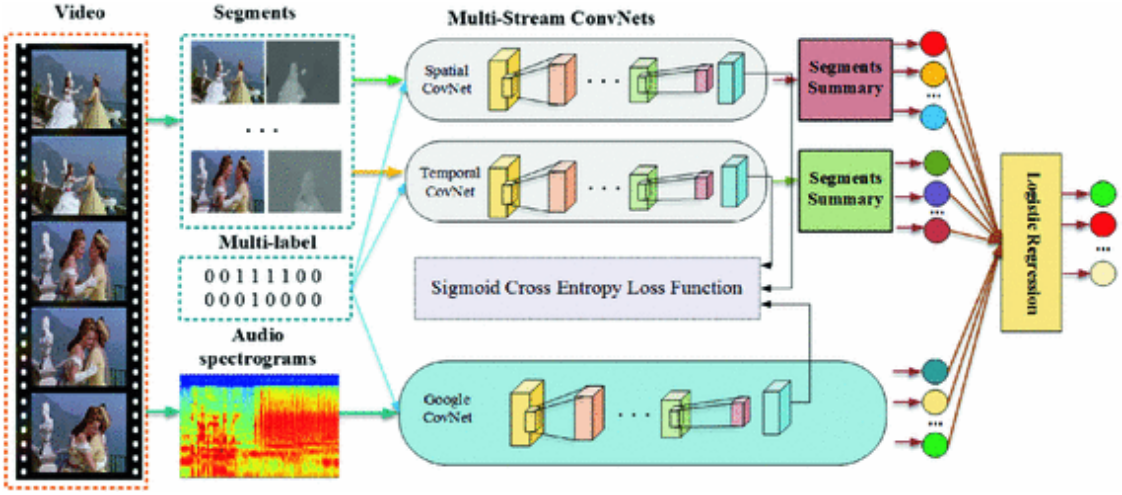


Figure 3.9. The Multi-stream Fusion Model [Lv et al., 2018] extracts multi-spectral information from global-scale video segments employing a TSN model [Wang et al., 2016b], while a GoogleNet [Szegedy et al., 2015] is used for spatial-temporal and audio spectrum features. The class scores obtained from each type of information are combined with a late fusion technique to obtain the final predictions.

fusion method, obtaining the final predictions, as shown by Figure 3.9. This work also provided a new dataset named Social Relation in Videos (SRiV), containing two splits of 8 relationship classes each, inspired by the Subjective Relations theory [Kiesler, 1983].

Aimar et al. [2019] presented the EgoSocialStyle, a dataset for social relation recognition from egocentric photostreams, including 5 classes inspired by the domain-based theory [Bugental, 2000]. The proposed model has a simple architecture, employing only personal-scale information from face and body frames, extracted using CNNs, and fed to an LSTM model [Hochreiter and Schmidhuber, 1997] for temporal reasoning, which generates the final features used for classification.

Liu et al. [2019] proposed a framework that exploits information from personal and global scales, with the addition of prior knowledge in the form of graphs representing the interactions between people and objects, similarly to Zhang et al. [2019]. From a set of frames sampled for each video, individual and object features are obtained, using the image regions detected by a Mask R-CNN [He et al., 2020], and extracting information with a ResNet [He et al., 2016]. Global spatial-temporal features are generated from the whole set of sampled frames, employing a TSN model [Wang et al., 2016b]. The gathered information is used to build three separated graph structures, which are fed to GCNs [Kipf and Welling, 2017] adapted to capture multi-scale information, producing the final feature vectors, which are fused and classified, as shown in Figure 3.10. The work also presented the Video-based Social Relation (ViSR) dataset, containing 8 relationship classes inspired by the domain-based theory [Bugental, 2000].

Finally, LV et al. [2019] extracts global features from video segments and optical flow, obtained from the sampled frames using CNNs. This information is directed to a

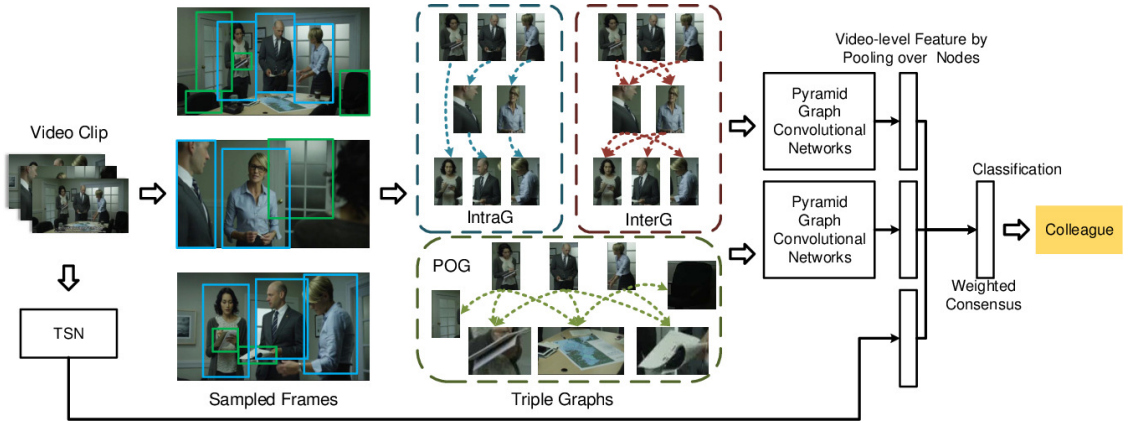


Figure 3.10. The Multi-scale Spatial-Temporal Reasoning model [Liu et al., 2019] extracts personal-scale information from body images and global attributes from objects, employing a ResNet [He et al., 2016], and a Mask R-CNN [He et al., 2020] for detection. This information is used to generate three separated graph structures representing interactions between persons and objects, which are fed to their respective GCN models [Kipf and Welling, 2017]. Global spatial-temporal features are also extracted with a TSN [Wang et al., 2016b], and fused with the features obtained from the graph structures, producing the final class scores.

module that combines LSTM cells [Hochreiter and Schmidhuber, 1997] with an attention mechanism to produce the final predictions, as shown in Figure 3.11.

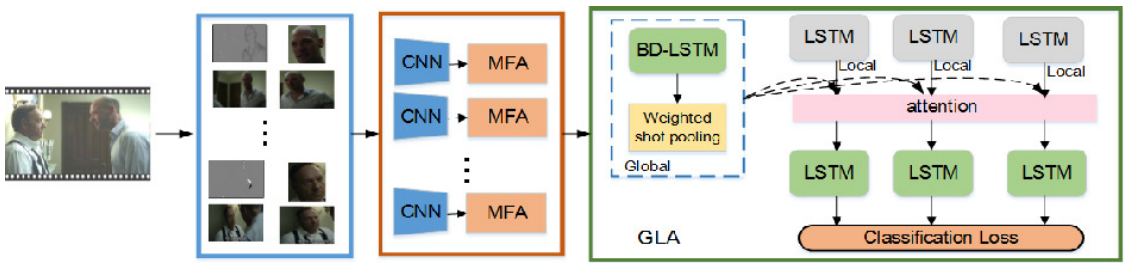


Figure 3.11. The model proposed by LV et al. [2019] captures global information from the sampled frames and their optical flow employing CNNs. This information is forwarded to an LSTM [Hochreiter and Schmidhuber, 1997] model, which also applies an attention mechanism to generate the final class scores.

The same authors also proposed a method [Dai et al., 2019] that extends Wang et al. [2018b] deep graph model to the temporal domain, but instead of employing information from personal and local-scale image regions, it learns spatial-temporal global information. This is done by extracting features from the set of sampled frames using convolutional networks and feeding this information to an LSTM model. The rest of the framework stands the same, as it can be seen from Figure 3.12.

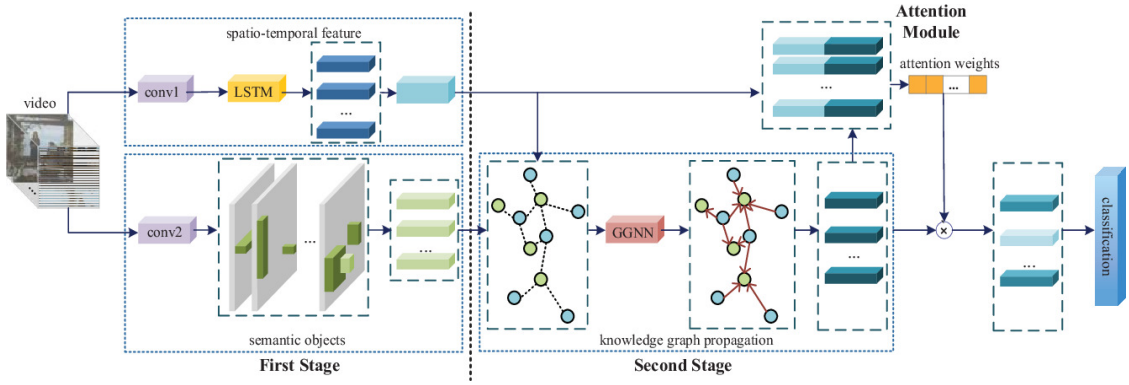


Figure 3.12. The Two Streams model [Dai et al., 2019] extracts global spatial-temporal features from a set of sampled frames and combines this information with object features using an attention mechanism, guided by a graph structure representing class co-occurrences for the whole dataset.

3.2 Social Relation Trait

Regarding social relation traits, the same type of analysis from the video-based methods is done in this section, focusing only on a brief description of methodological contributions, since this is not the same problem as the one considered in this work.

Zhang et al. [2015] was an early work that contributed to the area by presenting a new benchmark composed of 8 binary classes, inspired by the interpersonal circle theory [Kiesler, 1983]. It also proposed a siamese architecture capable of gathering information from personal-scale face images by extracting attributes including gender, expression, head pose, and age.

The method employs pre-trained CNN models combined with a bridging layer that leverages correspondences among the datasets used to learn attribute features by applying weak constraints derived from the association of face part appearances. Bounding boxes coordinates information is also extracted and concatenated with the pairwise feature vector to obtain the final classes scores, as shown in Figure 3.13.

Yan and Song [2019] used CNNs to extract data from personal and global scales. Individual features are obtained using face regions, while general information is extracted from the whole image. Both extraction models use ResNets [He et al., 2016] as their backbones, although the global network is combined with a semantic augmentation module that performs size, channel, and receptive field adjustments. The resulting feature vectors are concatenated and used to produce the final predictions, as shown in Figure 3.14.

Finally, Guo et al. [2019] employed a CNN model pre-trained for face recognition [Parkhi et al., 2015] to extract personal-scale features from face image regions, while incorporating spatial data from bounding box coordinates. The model also employs a ResNet [He et al., 2016] to obtain global information from the whole image. Both fea-

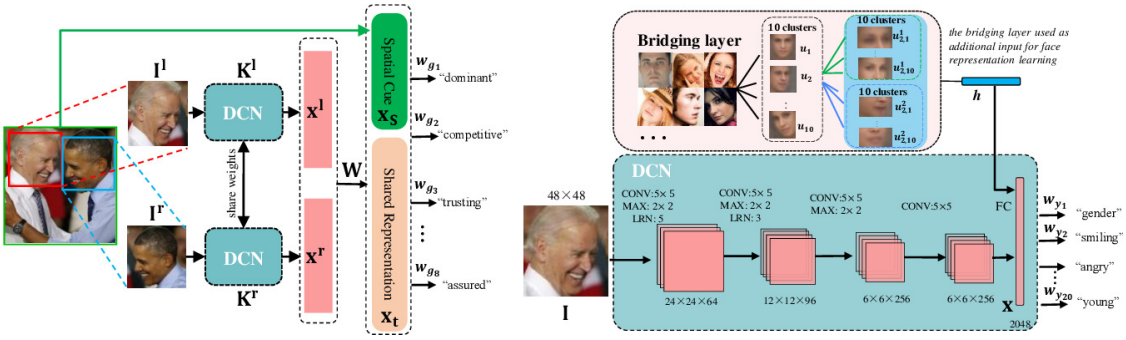


Figure 3.13. The siamese architecture proposed by Zhang et al. [2015], which extracts rich face representations from personal-scale face images using a convolutional model pre-trained on multiple attribute datasets. The features extracted from each person are concatenated, adding spatial cues obtained from bounding box coordinates, generating the final vector used for classification.

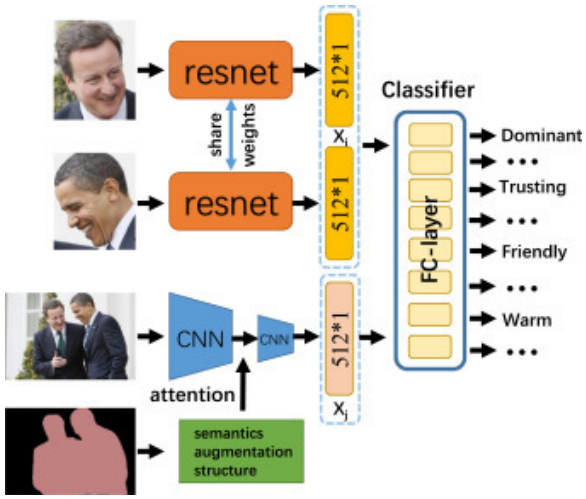


Figure 3.14. The Three Stream Network [Yan and Song, 2019] extracts personal-scale information from face images, which are combined with global-scale data obtained using the entire image. The three resulting feature vectors are concatenated and fed to a classifier.

ture sets are used to generate initial predictions, which are combined using a late fusion method, producing the final class scores, as illustrated in Figure 3.15.

3.3 Discussion

The previous sections presented a review of the existing works on social relation recognition and social relation trait from images and video data. Most of these methods employ personal and global-scale information, usually differing only on how this information is combined. The knowledge that can be obtained from other relationships in the image is totally ignored, with the exception of Li et al. [2020], which is able to capture

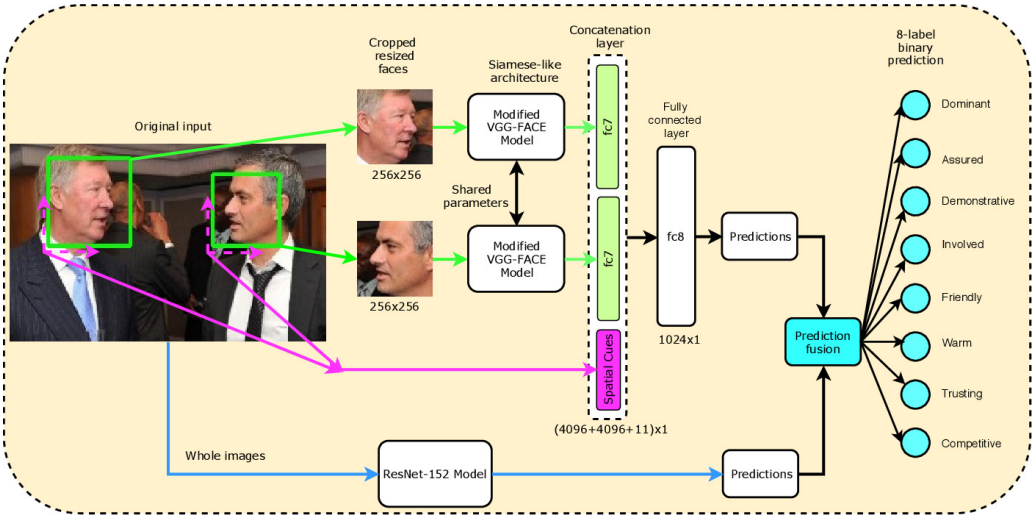


Figure 3.15. The architecture of the model proposed by Guo et al. [2019], combining personal-scale features employing face regions with global information from the whole image. The data obtained from both sources of information are used to generate intermediary predictions that are combined using late fusion to produce the final class scores.

limited personal-scale interdependencies. Considering attributes and constraints, all the works that handle them are capable of doing it only for one scale, and they are also unable to capture the dependencies between attributes from other relationships.

In short, all previous methods neglect somewhat relevant sources of information or destroy the original social relationship structure. This work introduces a graph representation capable of preserving this structure while also carrying and combining multi-scale learned features, prior knowledge, and other constraints. A deep graph model is also proposed to learn from this representation by exploiting its fundamental proprieties to aggregate the carried data into the high-level information necessary to identify social relationships.

3.3.1 Work Contextualization

A scheme including all previously examined works is presented in Table 3.1. They are organized according to the taxonomy proposed in Chapter 2, which considers their capacity to extract features directly from the image, capture interdependencies, and apply constraints in any form on each social scale. Finally, the last line contextualizes the contributions of this work.

Table 3.1. The reviewed works on social relationship recognition and relationship traits from image and video data, evaluated according to the taxonomy proposed in Chapter 2.

Task	**Extraction**			**Interdependencies**			**Constraints**		
Social Relation Traits	Personal	Local	Global	Personal	Local	Global	Personal	Local	Global
Zhang et al. [2015]	•						•	•	
Yan and Song [2019]	•		•						•
Guo et al. [2019]	•		•				•		
Social Relation Recognition									
Video									
Lv et al. [2018]			•						•
Aimar et al. [2019]	•								
Liu et al. [2019]	•		•	•		•			•
LV et al. [2019]			•						•
Dai et al. [2019]			•						•
Image									
Li et al. [2017]	•	•							•
Sun et al. [2017]	•						•		
Wang et al. [2018b]	•	•							•
Zhang et al. [2019]	•	•	•				•		•
Goel et al. [2019]							•	•	
Wang et al. [2020]							•		
Li et al. [2020]	•			•					
This Work	•	•	•	•	•	•	•	•	•

Chapter 4

Theoretical Framework

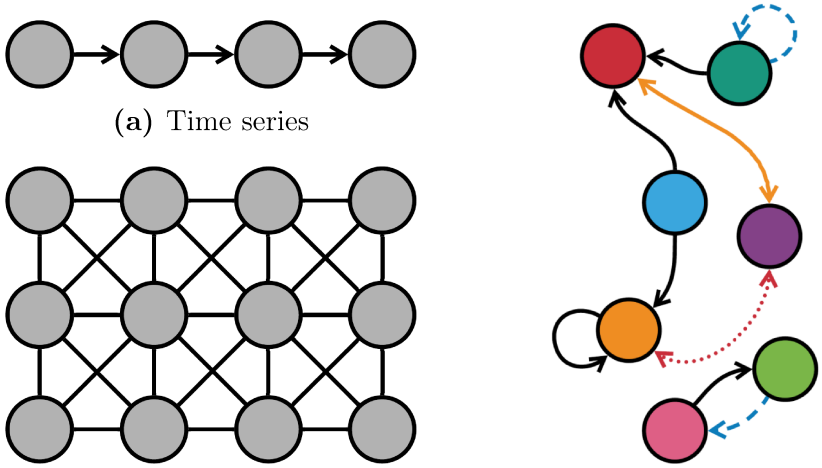
As mentioned in previous sections, this work proposes an approach to solve the social relation recognition problem by combining deep learning techniques and knowledge graph representations. This chapter presents a brief history of Graph Neural Networks (GNNs), along with the main aspects of each approach, their state-of-the-art methods, and applications. The provided information serves as a basis to describe the proposed methodology, which is introduced in the following chapters.

4.1 Graph Neural Networks

The advance of deep learning methods in recent years has heavily impacted areas such as image classification, video processing, and natural language understanding [Wu et al., 2021]. A key aspect of this success is that the data used for these tasks can usually be represented in the Euclidean space, which is not the case for graph-based data. However, recent research managed to improve the state-of-the-art significantly for Graph Neural Networks (GNNs), allowing an increasing adoption for multiple applications.

To comprehend the challenges in the development of GNNs, first, it is necessary to understand why previous deep models were so successful. Utilizing image data as an example, they have a well-defined structure that can be represented as a regular grid in the Euclidean space. A Convolutional Neural Network (CNN) is able to exploit proprieties from this structure such as shift-invariance, local connectivity, and compositionality to extract meaningful latent representations [Wu et al., 2021].

However, graph data is not structured in the same way, as shown in Figure 4.1, which means they do not have the same properties that can be explored in a similar form. Graphs may show irregular arrangements, with a variable number of unordered nodes, where each one can also have different numbers of neighbors. In some cases, they can present disconnected nodes, self-loops, and even carry information in their edges, increasing the difficulty of generalizing a convolution operation to the graph domain.



(a) Time series

(b) Image data

(c) Graph data

Figure 4.1. A visual representation showing information organized into different types of structures and how their proprieties impact the employed techniques. (a) Time series carrying sequential information, which can be exploited by considering previous states. (b) For image data, the nearby pixels are usually more correlated, and this trait is explored by CNNs. (c) A heterogeneous graph containing different types of nodes and edges, forming an irregular structure, which becomes a challenge for deep learning algorithms.

The first work to combine deep learning and graph representations was published by Sperduti and Starita [1997], motivating other initial studies [Gori et al., 2005; Scarselli et al., 2009; Gallicchio and Micheli, 2010] which further formalized and elaborated the concept of Graph Neural Networks. These works proposed approaches based on Recurrent Neural Networks (RNNs), which were used as iterative propagation models to learn node-level representations by aggregating neighborhood information.

In parallel, other methods inspired by the success of the CNNs were being developed, aiming to generalize the convolution operation for the graph domain by employing two distinct approaches. The spectral-based techniques started with Bruna et al. [2014], which presented a graph convolution derived from spectral graph theory, and subsequent publications [Henaff et al., 2015; Defferrard et al., 2016; Kipf and Welling, 2017; Levie et al., 2019] made further improvements with extensions and approximations on the graph convolution. The first spatial-based approach [Micheli, 2009] addressed graph mutual dependencies with nonrecursive layers and a message passing method inspired by recurrent GNNs, and later other spatial-based techniques [Niepert et al., 2016; Li et al., 2018; Duvenaud et al., 2015] emerged.

Recently, other neural network concepts were also extended to the graph domain, such as Graph Autoencoders [Cao et al., 2016; Wang et al., 2016a] and Spatial-Temporal Graph Neural networks [Jain et al., 2016; Seo et al., 2018]. With all this variety of methods, several studies [Gilmer et al., 2017; Wang et al., 2018a] suggested integrating all of them under a single general framework.

Finally, the recent advances in the area enabled a wide range of practical appli-

cations in computer vision, natural language processing, traffic forecasting, recommender systems, chemistry, biology, and others [Wu et al., 2021]. In the following sections, basic graph definitions are introduced, and some of the previously cited methods are further explored and exemplified with some of the most relevant works for each approach.

4.1.1 Definitions and Formalizations

In this section, some basic graph definitions and formalizations are introduced to serve as background for the technical explanations of each type of GNN in the following sections.

A graph can be represented as $G = (V, E)$, where V is the set of vertices or nodes, and E is the set of edges. Let $v_i \in V$ be a vertex, and $e_{i,j} = (v_i, v_j) \in E$ is the edge pointing from v_i to v_j. The neighborhood of a vertex v is defined as $N(v) = \{u \in V \mid (v, u) \in E\}$, and $A \in \mathbb{R}^{\mathcal{V}\times\mathcal{V}}$ is the adjacency matrix for the graph G, where $A_{i,j} = 1$ if $e_{i,j} \in E$ or $A_{i,j} = 0$ if $e_{i,j} \notin E$, while $\mathcal{V}$ is the total number of vertices.

Finally, vertices can carry attributes in the form of a feature matrix $F_v \in \mathbb{R}^{\mathcal{V}\times\mathcal{H}}$ where $f_v \in \mathbb{R}^{\mathcal{H}}$ represents the feature vector of the node v with dimension $\mathcal{H}$. Edges may also have attributes represented by the matrix $F_e \in \mathbb{R}^{\mathcal{E}\times\mathcal{H}}$ where $f^e_{v,u} \in \mathbb{R}^{\mathcal{H}}$ is the feature vector for the edge (v, u) with dimension $\mathcal{H}$, and $\mathcal{E}$ is the total number of edges.

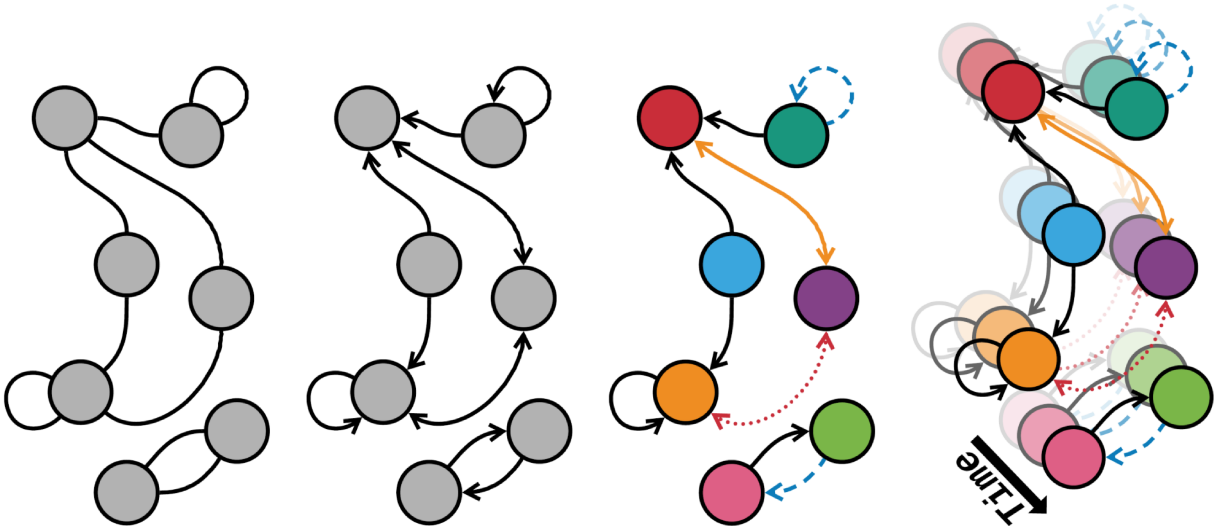
Depending on their characteristics, graphs can be classified in distinct ways, which can heavily influence the approaches used to extract their information. These types of graphs are described as:

Directed All the edges forming this kind of graph are directed from a vertex to another, as shown in Figure 4.2b. Undirected graphs (Figure 4.2a) are considered special cases where there is a counterpart, with reversed direction, for each edge in the graph. This means a graph is undirected if and only if the adjacency matrix is symmetric.

Heterogeneous Graphs of this type contain different kinds of nodes and edges, represented by distinct colors in Figure 4.2c.

Spatial-temporal This is a particular case of dynamic graphs that have a static structure where the nodes or edges attributes change dynamically over time (Figure 4.2d).

In most cases, GNNs take as input the graph structure represented by an adjacency matrix and the attribute feature vectors for nodes and edges. This information can be employed to perform three different levels of tasks:



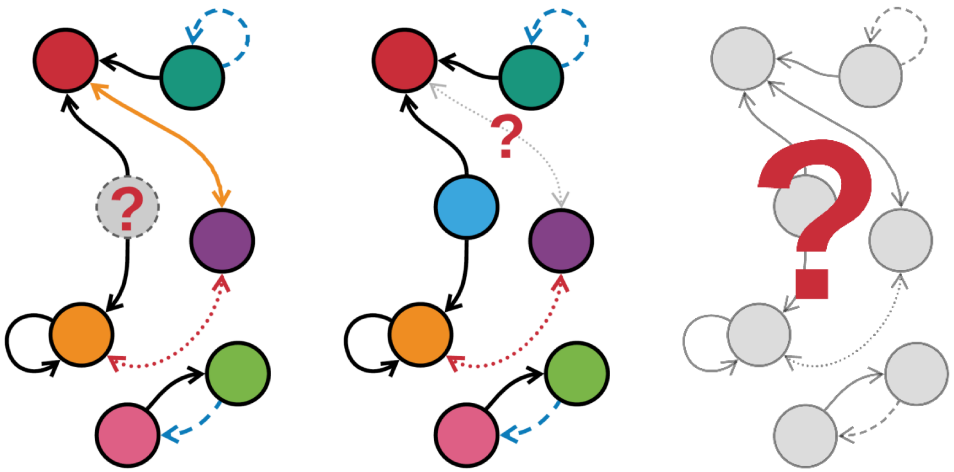
(a) Undirected **(b)** Directed **(c)** Heterogeneous **(d)** Spatial-Temporal

Figure 4.2. Representations of different types of graphs where the colors indicate distinct node and edge types. The arrows point the direction of the edges, and the transparent copies indicate the flow of time in the spatial-temporal graphs.

Node-level Extract high-level node representations, which can be applied to node regression and classification tasks (Figure 4.3a).

Edge-level Uses node representations to perform edge-related tasks such as edge classification and the prediction of the existence of edges between pairs of nodes, as shown in Figure 4.3b.

Graph-level Combines the final hidden states from all the nodes using a readout function to classify the graph as a whole (Figure 4.3c).



(a) Node-level **(b)** Edge-level **(c)** Graph-level

Figure 4.3. Representations of different levels of tasks involving graphs. The problem is usually approached by aggregating features from neighbor nodes and edges or even from the entire graph, depending on the type of task.

4.1.2 Recurrent Graph Neural Networks

Recurrent Graph Neural Networks (RecGNNs) apply the same set of parameters recurrently over the graph features to extract high-level representations. Initial research [Wu et al., 2021] mainly focused on acyclic graphs due to computational power restrictions. The first work capable of handling different types of graphs (e.g., acyclic, cyclic, directed, and undirected) was based on a mechanism designed to exchange node neighborhood information until an equilibrium point is reached [Scarselli et al., 2009]. The hidden state h_v^t for the node v at the time step t is obtained by

$$h_v^t = \sum_{u \in N(v)} \mathbf{f}(f_v, f_{v,u}^e, f_u, h_u^{t-1}), \tag{4.1}$$

where $\mathbf{f}(\cdot)$ is a parametric function, and in this case, it is learned by a neural network. When the convergence criterion is satisfied, the last hidden state is fed to a readout layer.

The Gated Graph Neural Network (GGNN) [Li et al., 2016] is one of the state-of-the-art contributions for the RecGNNs, employing a Gated Recurrent Unit (GRU) [Cho et al., 2014] as the propagation model and limiting the number of iterations, which ensures convergence without the need of constraining parameters. The hidden state update rule is defined as

$$h_v^t = \mathbf{GRU}(h_v^{t-1}, \sum_{u \in N(v)} W h_u^{t-1}), \tag{4.2}$$

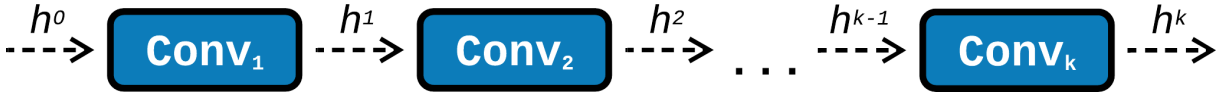
where W is a set of learnable parameters.

4.1.3 Convolutional Graph Neural Networks

Convolutional Graph Neural Networks (ConvGNNs) generalize the convolution operation for graph data by aggregating the node's own features with its neighbors' information. The main difference from RecGNNs is that each convolution layer runs a propagation step on the graph individually. In this sense, they can be stacked on top of each other, sending the message deeper into the graph for each layer, which also has a distinct set of learnable parameters instead of using the same parameters recurrently, as illustrated in Figure 4.4.

The layered format makes this type of GNN much more suitable to be employed with other neural network models, justifying the rapid growth in its application for various

(a) RecGNN



(b) ConvGNN

Figure 4.4. A visual representation of the general concepts behind the two main GNN approaches. (a) The RecGNN uses the same propagation model for each time step t. (b) For ConvGNN models, each convolutional layer propagates the information deeper into the graph while learning individual parameters.

tasks [Wu et al., 2021]. ConvGNNs works can be separated into two main approaches, spectral and spatial, which are defined by the method used to implement the convolution operation.

Spectral approaches define graph convolutions using filters from a graph signal processing perspective [Wu et al., 2021], where the operation can be interpreted as noise removal from graph signals. Assuming an undirected graph, the normalized graph Laplacian matrix L is defined as

$$L = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}, \tag{4.3}$$

where I is the identity matrix, D is the the diagonal matrix of node degrees, and $D_{i,i} = \sum_j (A_{i,j})$. The Laplacian can be factored as $L = U \Lambda U^T$, where $U = u_0, u_1, ..., u_{\mathcal{V}-1} \in \mathbb{R}^{\mathcal{V} \times \mathcal{V}}$ is the matrix of eigenvectors ordered by the eigenvalues, Λ is the diagonal matrix of eigenvalues where $\Lambda_{i,i} = \lambda_i$, and the normalized Laplacian form an orthonormal space, which means $U^T U = I$.

In graph signal processing, the Fourier transform to a signal x is defined as $\mathbf{F}(x) = U^T x$, projecting the input signal to the orthonormal space and the inverse transform is obtained by $\mathbf{F}^{-1}(\hat{x}) = U\hat{x}$, where $\hat{x}$ represents the transformed signal. In this way, the graph convolution of the input signal x with a filter $g \in \mathbb{R}^{\mathcal{V}}$ is define as

$$x *_G g = \mathbf{F}^{-1}(\mathbf{F}(x) \odot \mathbf{F}(g)) = U(U^T x \odot U^T g), \tag{4.4}$$

where $\odot$ denotes the element-wise product. If we write a filter as $g_\theta = diag(U^T g)$, the spectral convolution operation can be simplified as

$$x *_G g_\theta = U g_\theta U^T x. \tag{4.5}$$

All spectral-based methods are based on this definition, and their main difference is the choice of the filter g_θ. For example, the Chebyshev Spectral CNN (ChebNet) [Defferrard et al., 2016] uses the Chebyshev polynomials of the diagonal matrix of eigenvalues

to approximate the filter g_θ, with the resulting convolution operation being calculated as

$$x *_G g_\theta = U \left(\sum_{i=0}^{K} \theta_i T_i \left(\frac{2\Lambda}{\lambda_{max} - I} \right) \right) U^T x. \tag{4.6}$$

Finally, the Graph Convolutional Network (GCN) [Kipf and Welling, 2017] is a spectral-based approach that received a lot of attention recently. The method proposed a first-order approximation of the ChebNet, assuming $K = 1$ and $\lambda_{max} = 2$, obtaining

$$x *_G g_\theta = \theta_0 x - \theta_1 D^{-\frac{1}{2}} A D^{-\frac{1}{2}} x. \tag{4.7}$$

In order to restrain the number of parameters, avoiding overfit, the GCN assumes $\theta = \theta_0 = -\theta_1$, obtaining the final definition of the graph convolution as

$$x *_G g_\theta = \theta(I + D^{-\frac{1}{2}} A D^{-\frac{1}{2}}) x. \tag{4.8}$$

Spatial methods recycle RecGNN ideas to define convolutions by information propagation, increasing their efficiency, flexibility, and generality when compared to spectral techniques, which resulted in rapid growth in their adoption [Wu et al., 2021]. The Neural Network for Graphs (NN4G) [Niepert et al., 2016] is the first approach to use a spatial-based convolution by directly summing neighborhood information and applying residual connections. This convolution operation can be defined as

$$h_v^k = \mathbf{f}(W^{(k)^T} f_v + \sum_{i=1}^{k-1} \sum_{u \in N(v)} Z^{(k)^T} h_u^{k-1}), \tag{4.9}$$

where W^k and Z^k are the learnable parameters of the layer k, and $\mathbf{f}(\cdot)$ is an activation function.

Graph Attention Networks (GAT) [Veličković et al., 2018] assume neighbor node contributions are not the same, adopting an attention mechanism to learn relative weights between two connected nodes. The GAT convolution is obtained by

$$h_v^k = \sigma(\sum_{u \in N(v) \cup v} \alpha_{v,u}^k W^k h_u^{k-1}), \tag{4.10}$$

where W^k is the set of learnable parameters of the layer k and σ is the sigmoid activation function. The attention weights $\alpha_{v,u}^k$ between the nodes v and u are can be calculated by

$$\alpha_{v,u}^k = \mathbf{softmax}(\mathbf{g}(a^T [W^k h_v^{k-1} || W^k h_u^{k-1}])), \tag{4.11}$$

where $\mathbf{g}(\cdot)$ is the LeakyReLU activation, and the softmax function ensures the attention weights sum up to one over all neighbor nodes.

Although spectral models have a solid theoretical foundation in graph signal processing, spatial approaches are preferred due to efficiency, generality, and flexibility advantages [Wu et al., 2021]. This happens because spectral methods have to execute eigenvector calculation while also handling the whole graph, which makes them more computationally expensive and susceptible to scalability problems. Spatial models solve this problem by performing convolutions directly in the graph domain using information propagation, also allowing node batching.

Finally, spectral methods assume fixed and undirected graphs, leading to generalization problems since any perturbations would result in a change of eigenbasis. On the other hand, spatial models calculate graph convolutions locally on each node, allowing parameters to be easily shared across different locations. They also can handle multiple types of structures, including edge inputs, directed, signed, and heterogeneous graphs, since all this information can be incorporated into the aggregation function [Wu et al., 2021].

4.1.4 General Frameworks

With the increasing number of new graph neural network methods, some works proposed general frameworks, aiming to integrate distinct approaches under a single methodology. The Message Passing Neural Networks (MPNN) framework [Gilmer et al., 2017] abstracts commonalities between several methods such as spectral approaches [Kipf and Welling, 2017; Bruna et al., 2014; Defferrard et al., 2016], spatial-based models [Duvenaud et al., 2015] and even RecGNNs [Li et al., 2016]. The framework is composed of two main steps: message passing and readout. The first step runs for $\mathcal{T}$ times and is defined by the message $\mathbf{M}_t$ and the update $\mathbf{U}_t$ functions at the time step t obtained by

$$m_v^{t+1} = \sum_{u \in N(v)} \mathbf{M}_t(h_v^t, h_u^t, e_{v,u}), \tag{4.12}$$

$$h_v^{t+1} = \mathbf{U}_t(h_v^t, m_v^{t+1}). \tag{4.13}$$

After the message propagation process, the readout step computes the final representation for the whole graph using the function $\mathbf{R}$, defined as

$$\hat{y} = \mathbf{R}(h_v^{\mathcal{T}} \mid v \in G). \tag{4.14}$$

The Non-Local Neural Networks (NLNN) [Wang et al., 2018a] is a general frame-

work with the purpose of capturing long-range dependencies using deep models. The method generalizes the non-local operation [Buades et al., 2005] from computer vision to graph data, computing a weighted sum of the features at all positions, which can be in space, time or space-time. In this sense, the NLNN can be seen as a unification of various self-attention methods [Zhou et al., 2020]. The generic non-local operation is defined as

$$h_i^* = \frac{1}{\mathcal{N}(h)} \sum_{\forall j} \mathbf{f}(h_i, h_j)\mathbf{g}(h_j), \tag{4.15}$$

where i is the index of an output position, j is the index enumerating all possible positions, $\mathbf{f}(\cdot)$ computes a scalar representing the relation between the parameters, $\mathbf{g}(\cdot)$ denotes a transformation of the input, and $\frac{1}{\mathcal{N}(h)}$ is a normalization factor. There are several options for these operations, such as the gaussian function, dot product, and concatenation.

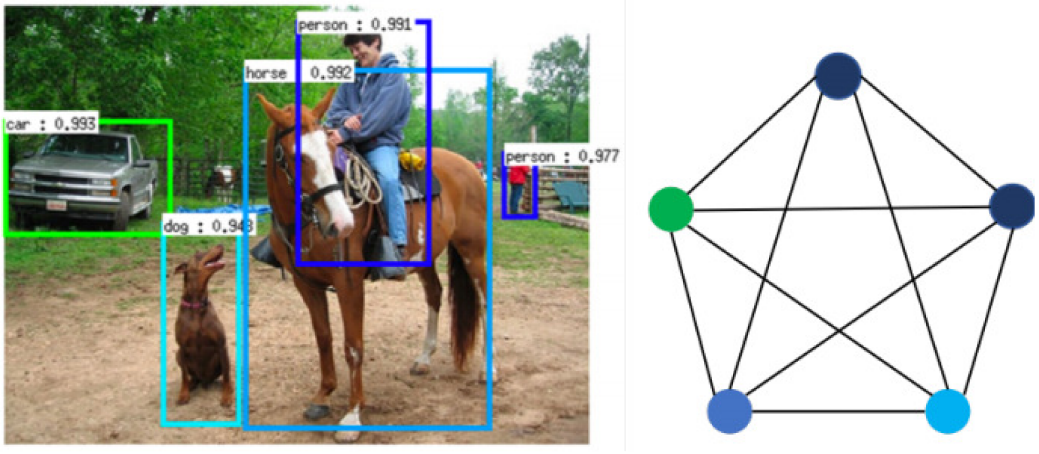
4.1.5 Applications

Graph Neural Networks have an extensive set of applications, since they are able to extract information from any data that can be represented as graphs. Usually, graph data can be separated into two categories based on how it is arranged. For structural information, there is an explicit relational structure, for example, physical systems, molecular structures, and knowledge graphs [Zhou et al., 2020]. However, in non-structural scenarios, this relational structure is not explicit, such as for images and text. Some examples of graph applications in different circumstances are shown in Figure 4.5.

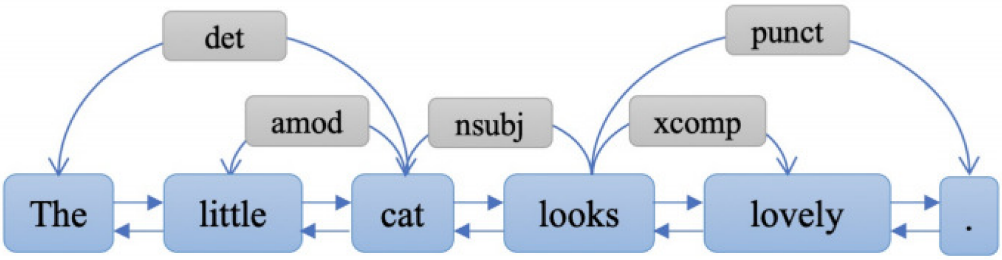
In structured scenarios, the main applications are in biology [Duvenaud et al., 2015], chemistry [Gilmer et al., 2017], traffic forecasting [Li et al., 2018] and recommender systems [Ying et al., 2018]. However, it is for non-structured scenarios where the advantage of representing information as graphs has opened a new set of possibilities, including well-known computer vision and language processing problems [Zhou et al., 2020].

For example, in group action recognition, Wu et al. [2019] generates a graph encoding appearance and positioning data. This strategy can also be helpful for visual question answering, where Teney et al. [2017] constructed graphs representing scene objects and their spatial arrangement. Graphs can also be applied to other types of tasks such as feature learning, as done by Meng et al. [2018] for relative attribute learning, and Guo et al. [2020] for feature selection in a group emotion and event recognition context.

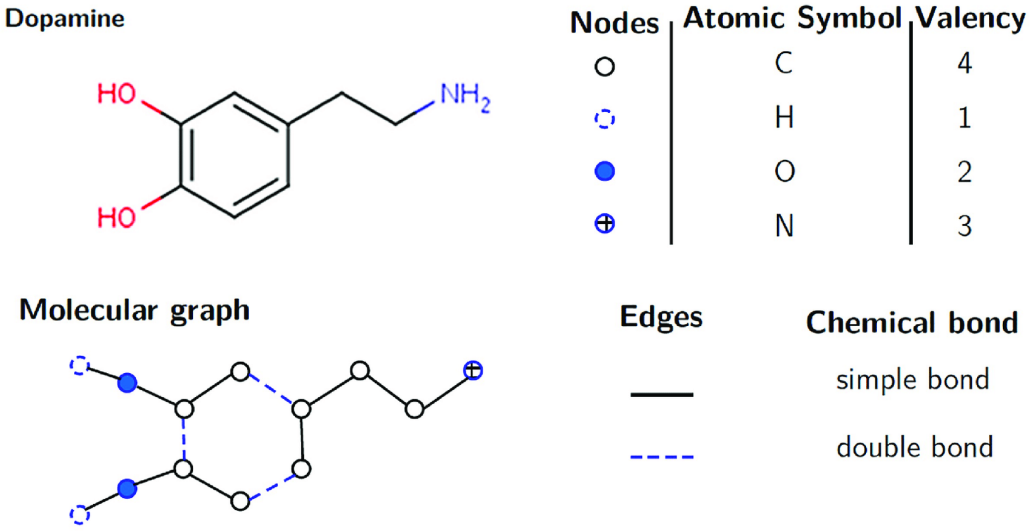
In this work, a knowledge graph is used to represent the relationships depicted in an image, preserving their original structure and capturing the dependencies between learned features and attributes from multiple scales. Next, a GNN model is proposed to



(a) Computer Vision



(b) Language Processing



(c) Chemistry

Figure 4.5. Some practical applications for GNNs. (a) Objects and their relationships can be represented as graphs for computer vision problems [Zhou et al., 2020]. (b) For language processing tasks, the textual structure can be designed as a connected graph [Zhou et al., 2020]. (c) Molecules can be directly translated into graphs for chemistry and biology-related tasks [Ilemo et al., 2019].

extract the information from the graph, obtaining high-level features that can be used to classify each relationship.

Chapter 5

Methodology

In this chapter, the original problem formulation is revisited and rearranged to fit the a image-based paradigm instead of the former pairwise approach. Next, the methodology is explained along with its design choices and technical aspects, which are further formalized and exemplified. An overview of the proposed framework is shown in Figure 5.1, including image processing, feature extraction, graph construction, relation reasoning, and classification.

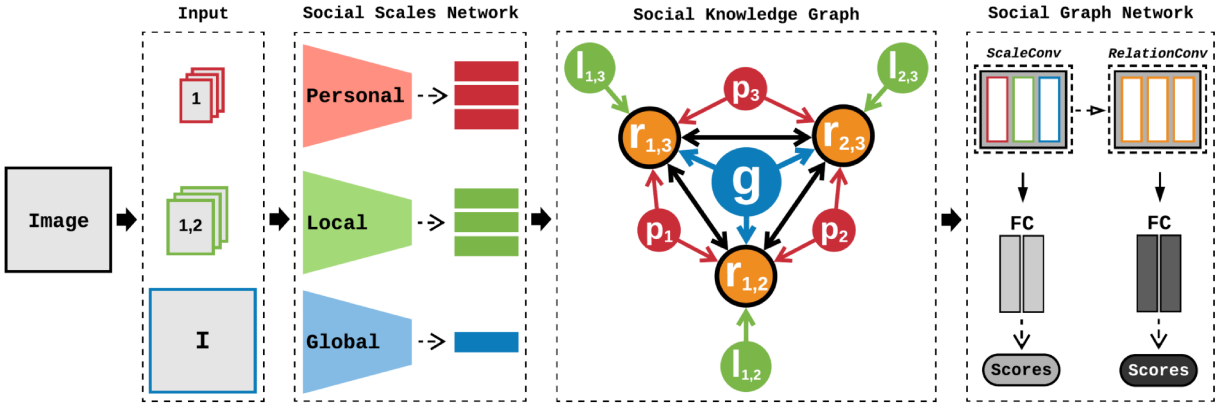


Figure 5.1. An overview of the implemented framework, employing CNNs as extraction backbones within the Social Scales Network (SSN) to obtain features from distinct regions of the input image. This information is used to initialize the nodes from the Social Knowledge Graph (SKG) representing the social structure for the whole image. Finally, the proposed Social Graph Network (SGN) performs reasoning on this graph, generating an updated representation for each social relationship, which is used for the classification.

The approach extracts information from three distinct social perspectives in a given image: personal, local, and global, which are referred to as social scales and denoted by p, l, and g, respectively. They are sources of individual, relative, and general features, offering essential complementary data, as described in Section 2.1. The model is also capable of incorporating prior knowledge in the form of attributes to each one of these scales, indicating the most important traits to consider. They act as constraints, guiding the network in the direction of identifying social relationships in a similar way as humans do, which is a fundamental behavior for social relation tasks, as explained in Section 2.3.

All the extracted information is carried by a graph indicating how to combine it, modeling dependencies between scales, attributes, and relationships. Unlike other

techniques, the proposed graph representation retains the social structure connecting all relationships within an image, preserving their intradependencies and interdependencies. This allows the model to consider other relation pairs and their multi-scale attributes during the reasoning process, which are essential sources of information, since they can be strongly correlated, as shown in Sections 2.2 and 2.3.

The implementation combines convolutional and graph neural networks to extract features directly from the input image, building the social graph representation and performing reasoning over it in an end-to-end framework. The described process is composed of three modules, namely, Social Scales Network (SSN), Social Knowledge Graph (SKG), and Social Graph Network (SGN), which are further detailed in the following sections.

5.1 Problem Formulation

This section presents an analysis of the original problem definition provided by other works. The obtained conclusions are used as basis for the development of the proposed methodology, which reformulates the former approach to reflect the interdependencies between relationships in the same image.

Given an input image $\mathbf{I}$ depicting $\mathcal{P}$ individuals, a set of features $F = \{f_i \mid i = 1, 2, ..., \mathcal{P}\}$, where f_i is the information corresponding to the i-th person, and also considering the queries

$$Q = \{q_{i,j} \mid i = 1, 2, ..., \mathcal{P}, j = 1, 2, ..., \mathcal{P}, i \neq j\}, \tag{5.1}$$

where $q_{i,j}$ is the relationship involving the persons i and j, the social relation recognition problem was defined by Goel et al. [2019] as finding the optimal value

$$Q^* = \arg\max_{Q} Pr(Q \mid \mathbf{I}, F), \tag{5.2}$$

where

$$Pr(Q \mid \mathbf{I}, F) = \prod_{i=1}^{\mathcal{P}} \prod_{\substack{j=1 \\ j \neq i}}^{\mathcal{P}} Pr(q_{i,j} \mid \mathbf{I}, f_i, f_j). \tag{5.3}$$

However, as it can be seen from Equation 5.3, this type of approach assumes the relationships depicted in the input image are independent, which is not valid, as explained in Section 2.2. By considering this paradigm, earlier works optimize the objective function for each pair separately, and therefore the value obtained for Equation 5.2 is not optimal.

In this work, the relationships are treated as dependent events, and the social

relation recognition problem is solved by considering information from the entire image, optimizing the model with specific features and attributes from all individuals and their relationships jointly. To achieve this, the original metadata and ground-truth labels from the employed datasets were adapted to fit an image-based approach instead of the pair-based one used in previous works. More detailed information on this process is provided in Section 6.3.

Additionally, it is important to note that a pairwise approach also exacerbates the problems related to the local and global scales, as mentioned in Section 2.1. This happens because, in some cases, similar local region patches and global images will continually be fed to the model for every relationship in the image, but with different corresponding labels, generating severe inconsistencies and hindering the features learned by these models.

This problem is solved in the proposed method because every global image patch is presented to the model only once per image, and the extracted features are not directly classified. Instead, they are combined with more specific information provided by other scales, correctly identifying the pair of individuals participating in each relationship, which generates adequate feedback to the model during the training process.

Finally, the concept of social neighbors is introduced, allowing the model to filter the information from other relations. This helps to reduce the noise generated by considering all the relationships within an image together, since some of them may not be correlated or, in some circumstances, they can be missing from the dataset annotations.

5.2 Social Scales Network

The purpose of this step is to extract visual information directly from the input, learning features that will be used to initialize the hidden states of the nodes from the graph representing the social structure in the image. This data serves as a starting point for the relation reasoning model, which exploits the graph structure to aggregate the extracted information, generating high-level features, which will be finally employed in the final classification.

This module is composed of two sub-modules, as illustrated in Figure 5.2, that are in charge of pre-processing the input images and performing feature extraction. Social-scales data is obtained directly from image patches using Convolutional Neural Networks (CNNs) to gather information specifically from each scale, which means three distinct convolutional backbone models are employed. The first one extracts individual data from personal-scale body images, the second obtain local-scale relative features for each

relationship by combining the images of both participating individuals, and the last one is responsible for extracting global-scale general information from the image as a whole.

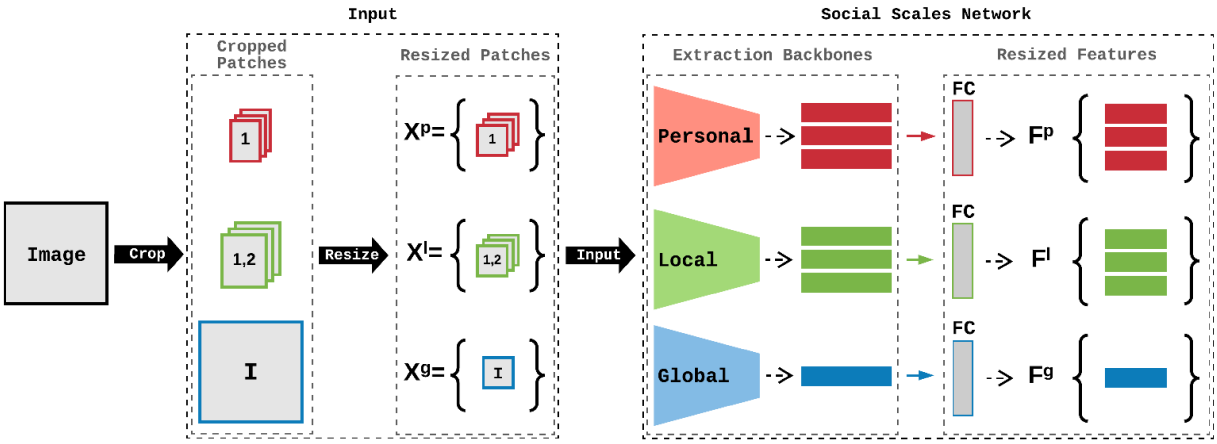


Figure 5.2. A detailed representation of the module, which is composed of two main stages. The first one processes the input image, cropping and resizing social-scale region patches. The second employs three distinct CNN models to extract information from each scale, receiving the processed image patches and outputting the corresponding feature vectors.

As stated in Section 2.1, visual information can exhibit different meanings according to context. For this reason, the objective of each backbone model is to specialize in the visual features that are most relevant to their respective social scales. This is achieved by cropping the regions corresponding to each scale from the input image, as shown in Figure 5.3, and feeding these patches to their corresponding backbone models. The image-processing task is performed by an initial sub-module, which receives the input image and outputs cropped and resized image patches to a second sub-module named Social Scales Network (SSN), which contains the backbone convolutional models.

(a) Input image **(b)** Resized image patches

Figure 5.3. An example showing (b) the sets of resized region patches X^p, X^l and X^g, generated from (a) the given input image.

Each backbone model is expected to extract different granularities of detail, since these networks have the same architecture, and the input patches are downscaled to the

same dimensions. In other words, body images will carry fine-grained personal features, and as the scale goes up, the respective networks receive images containing coarser-grained traits and additional context-based information, up to global level features, as shown in Figure 5.3b.

More specifically, the input sub-module receives an image $\mathbf{I}$ depicting a number $\mathcal{R}$ of relations between $\mathcal{P}$ individuals, along with a set $B = \{b_i \mid i = 1, 2, ..., \mathcal{P}\}$ containing their bounding boxes coordinates, and the set $R = \{r_{i,j} \mid i = 1, 2, ..., \mathcal{P}, j = 1, 2, ..., \mathcal{P}, i \neq j\}$ with size $\mathcal{R}$, defining all the relation pairs in $\mathbf{I}$, where $r_{i,j}$ indicates there is a relationship between persons i and j.

Initially, the bounding boxes areas for all persons are cropped, generating a set of personal-scale input images $X^p = \{x_i \mid (\forall b_i \in B)[x_i = \mathbf{crop}(b_i)]\}$. Next, the set of local-scale input images $X^l = \{x_{i,j} \mid (\forall (r_{i,j} \in R)[x_{i,j} = \mathbf{crop}(r_{i,j})]\}$ is formed, where each element is obtained by cropping the regions defined by the smallest area from $\mathbf{I}$ containing the bounding boxes b_i and b_j of the persons participating in the relationship $r_{i,j}$. Finally, the whole image $\mathbf{I}$ is also used to represent the global scope, denoted by X^g.

The obtained image sets X^p, X^l and X^g are resized to the adequate input dimensions and fed to their respective social-scale backbone in the next sub-module. Each of these models have a Fully Connected (FC) layer, followed by the **ReLU** activation attached to their last feature extraction layer, with the purpose of resizing their outputs to the hidden state dimension $\mathcal{H}$ of the proposed graph model. After the described processes, the module outputs the sets $F^p \in \mathbb{R}^{\mathcal{P} \times \mathcal{H}}$, $F^l \in \mathbb{R}^{\mathcal{R} \times \mathcal{H}}$ and $F^g \in \mathbb{R}^{\mathcal{H}}$ of *personal*, *local* and *global* scale features.

5.3 Social Knowledge Graph

The objective of this second module is to build a representation that embodies all the information extracted from the image in the previous step. This representation will be used for the relation reasoning process in the last stage, and therefore it needs to preserve the social structure while also being able to apply constraints on how all these distinct types of information should be associated. This is achieved by the Social Knowledge Graph (SKG), a directed heterogeneous graph formed by four different node types: relationship, personal, local, global, and a corresponding edge type for each one, as depicted in Figure 5.4.

The SKG has an intuitive construction process, which is described in three steps for simplification purposes. The first one elaborates on the underlying concepts, also defining relationship nodes, which form the basis for this graph structure. The second

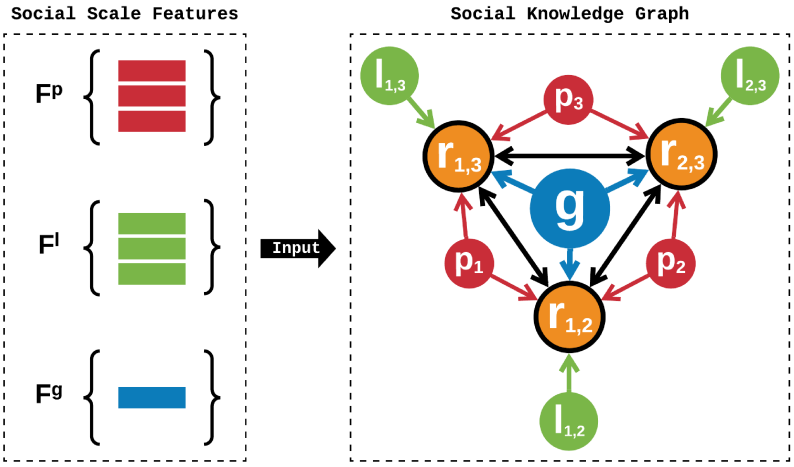
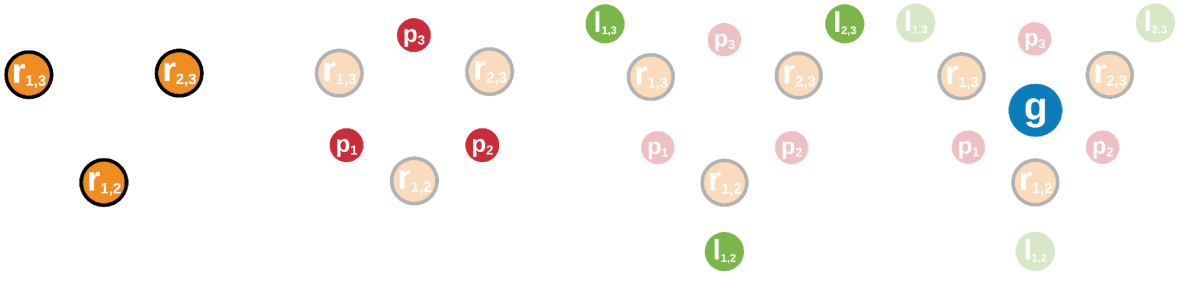


Figure 5.4. A visual representation of the SKG generated with the features extracted from the input image (Figure 5.3a) by the previous module.

part describes how to build social-scale nodes and their initialization, while the last step describes edge variants, how to establish connections between each type of node, and their role in the proposed structure. Finally, the design choices and the inner workings of this graph representation are explained, introducing the concepts of relationships structure and social neighbors.

Since the purpose of the model is to classify relationships, the building procedure starts with the creation of such node types to represent them, forming the basis of the graph structure, and their main function is to hold the aggregated features that will be employed in the final classification. These features are generated after the relation reasoning process, which considers the information carried by the node neighborhood and embedded in the graph structure. More specifically, the construction of the SKG begins with the creation of a relation node $\mathbf{r}_{i,j}$ for each social relationship $r_{i,j} \in R$. Additional information, such as social-scale features are assembled around their corresponding relation nodes, forming subgraph structures, as shown by each step in Figure 5.5.



(a) Relation **(b)** Personal **(c)** Local **(d)** Global

Figure 5.5. Step-by-step illustration of the node generation process for the SKG, starting with (a) the relation nodes representing every relationship. During the following steps, (b) personal, (c) local, and (d) global nodes are added, carrying the features extracted by their respective social-scale backbone models.

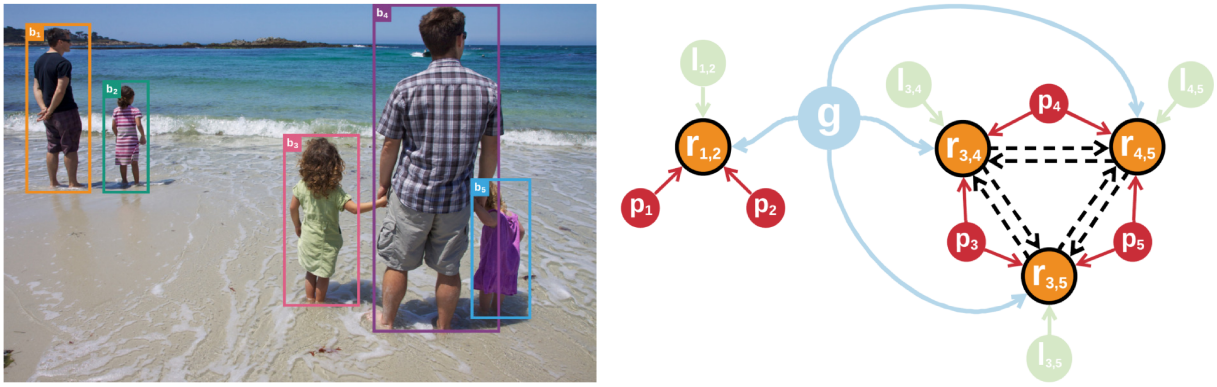
The SKG defines node types for each social scale, which are responsible for holding

the features extracted by the previous module. In this sense, a node of the corresponding type is generated for every distinct feature vector, and together they form a subgraph representing each scale. This means that a personal-type node $\mathbf{p}_i$ is created for every individual $i = 1, 2, ..., \mathcal{P}$ in the input image, receiving their respective features $f_i \in F^p$ as their initial hidden state h_i^p. For each relation pair $r_{i,j} \in R$, a node $\mathbf{l}_{i,j}$ of personal-type is built and its initial hidden state $h_{i,j}^l$ is fed with the corresponding features $f_{i,j} \in F^l$ obtained from pairwise image patches. Finally, a single global-type node $\mathbf{g}$ is created to carry the features $f \in F^g$ in its initial hidden state h^g, representing the input image as a whole.

This next step consists of defining associations between these generated nodes, establishing rules on how all the information extracted from the input image should be combined to classify each relationship. The linking process is based on the source of the features each one carries, beginning with social-scale nodes, which are connected to their respective relation nodes by social-scale edges of the corresponding types. More precisely, personal nodes are linked by directed personal edges $(\mathbf{p}_i, \mathbf{r}_{i,j})^p$ to the relation nodes in which their respective individual i participate, and local nodes have a direct correspondence with their relation nodes through directed local edges $(\mathbf{l}_{i,j}, \mathbf{r}_{i,j})^l$. Finally, since global features can affect all the social relations depicted in the image $\mathbf{I}$, all relation nodes are linked to the global node by directed global edges $(\mathbf{g}, \mathbf{r}_{i,j})^g$.

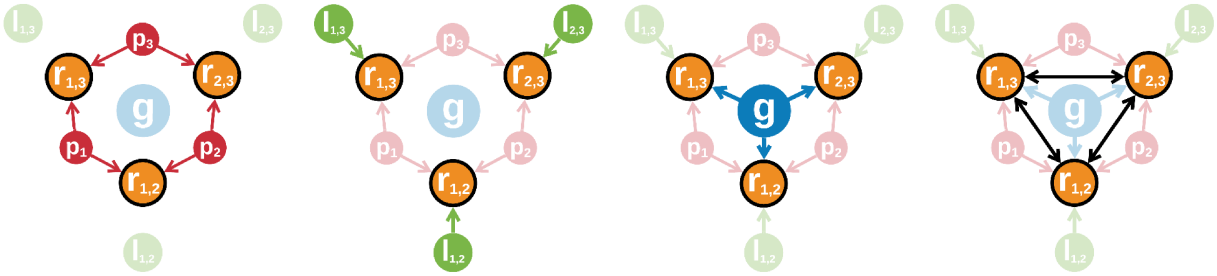
At this point, the graph contains nodes representing all the social relations depicted in the image, connected to their respective social-scale nodes, offering multi-scale information. However, this structure is treating relationships independently, and to fix this issue, a pair of edges is inserted in between all the relation nodes that have persons in common. This is the concept of *social neighbors*, represented in Figure 5.6, and through these connections, the model will be able to exchange information among distinct relationships of the same individual, learning the dependencies existing between some classes of relations, as described in Section 2.2. The set of social neighbors of the node v is represented by $S(v)$, and the directed relationship edges $(\mathbf{r}_{i,j}, S(\mathbf{r}_{i,j}))^r$ will form a two way link between every pair of social neighbors when applied to all relation nodes, allowing these nodes to send and receive information evenly.

An image can contain a total of $\mathbf{C}(\mathcal{P}, 2)$ social relationships, according to the number of depicted persons, as mentioned in Section 1.3. However, many samples in the benchmark datasets do not have ground-truth annotations for all possible pairs. This means that inserting relation edges between these nodes could generate noise instead of helping the classification. For this reason, relationships are connected only to their social neighbors, avoiding the addition of unnecessary information while also supporting the reasoning model to identify instances that are not correlated using their features. The entire process of attaching edges to the graph is shown in Figure 5.7, and after this final step, the construction of the default SKG version is completed.



(a) Example image **(b)** Corresponding SKG

Figure 5.6. An illustration of the *social neighbors* concept, where relation nodes connected to the same person node are linked directly by edges pointing in both directions. For simplification purposes, they can also be represented as a bidirectional edge.



(a) Personal **(b)** Local **(c)** Global **(d)** Relation

Figure 5.7. Step-by-step illustration of the edge addition process to the SKG, starting from social-scale nodes, which are linked to their respective relation nodes by an edge of the same type. (a) Personal edges connect each person node to the relationship in which they are involved, then (b) local edges link local context nodes to the graph, which have a direct correspondence with relation nodes, and (c) global edges connect the global node to all relationships. Finally, (d) relation edges are inserted between relation nodes and their social neighbors, concluding the construction of the graph.

As stated in previous sections, to improve the performance of social relation recognition methods, it is crucial to develop approaches that are more similar to how humans perceive social relationships. This is the main purpose of the SKG, and it is achieved by preserving the original structure of the relationships, which guides the flow of information during the relation reasoning process. More specifically, the graph carries multi-scale features and clearly defines in which relationships every individual takes part, allowing the exchange of information between them. This structure is destroyed by other methods when they crop these persons from the input image and try to infer their relationship independently.

Another important advantage of the proposed SKG is the ability to batch multiple samples within a single graph, where each connected component represents an image, preventing the interference of unrelated information. The reasoning process can be applied to all these samples simultaneously without the need of padding the graph with blank

nodes, like in previous approaches. This allows the model to learn on bigger structures, accelerating and adding more consistency to the entire process.

5.3.1 Attribute Nodes

This sub-module is in charge of adding attribute nodes to the SKG, generating an extended version of the graph. The features these nodes carry are extracted employing a pre-trained model for each attribute type, using as input the same image patches from which the social-scale features were extracted. These attributes add extra information into the graph structure, acting as constraints and allowing the model to learn the dependencies between them.

As mentioned in Section 1.3, early methods suffer from a recurrent problem where the developed models learn incorrect associations from the training data, leading to generalization issues. This work presents the hypothesis that those problems emerge because the decision-making process does not reflect how humans interpret social relationships, which is a fundamental behavior for these models since these relations are based on human perception.

The way how humans differentiate between social relationships is heavily influenced by appearance attributes such as age, gender, clothing, emotion, and body positioning, as described in Section 2.3. Additionally, previous research [Sun et al., 2017; Wang et al., 2020] quantified the impact of these attributes for social relation recognition, allowing the selection of the most meaningful traits. The proposed methodology applies this knowledge with the purpose of guiding the reasoning process to become more similar to how humans perceive social relationships.

In this sense, the goal of attribute nodes is to support the model in considering these essential aspects. They act as constraints to the decision-making process, since they are not learned directly from the input image, but instead, they are provided by models pre-trained on each specific trait. These types of nodes are attached to their respective scale nodes carrying the image patch features they were extracted from, and for this reason, they are considered social-scale attributes. For example, appearance traits such as age and gender can be obtained from the person image patches X^p, and hence they are personal attributes. From the local-scale pair images X^l, it is possible to extract features for activity or group emotion, which are considered local traits. Finally, it is possible to detect objects within the whole image $\mathbf{I}$, and their features are attached to the graph as global-scale attributes.

The insertion of attribute nodes to the graph begins by choosing which traits and

pre-trained models will be employed for each scale. Next, this information is extracted using the same image patches generated from their respective social scales by the input sub-module. Finally, the extracted features need to be resized to the hidden state dimension, since the extraction models probably output different vector sizes, and the relation reasoning model works with a fixed hidden state dimension.

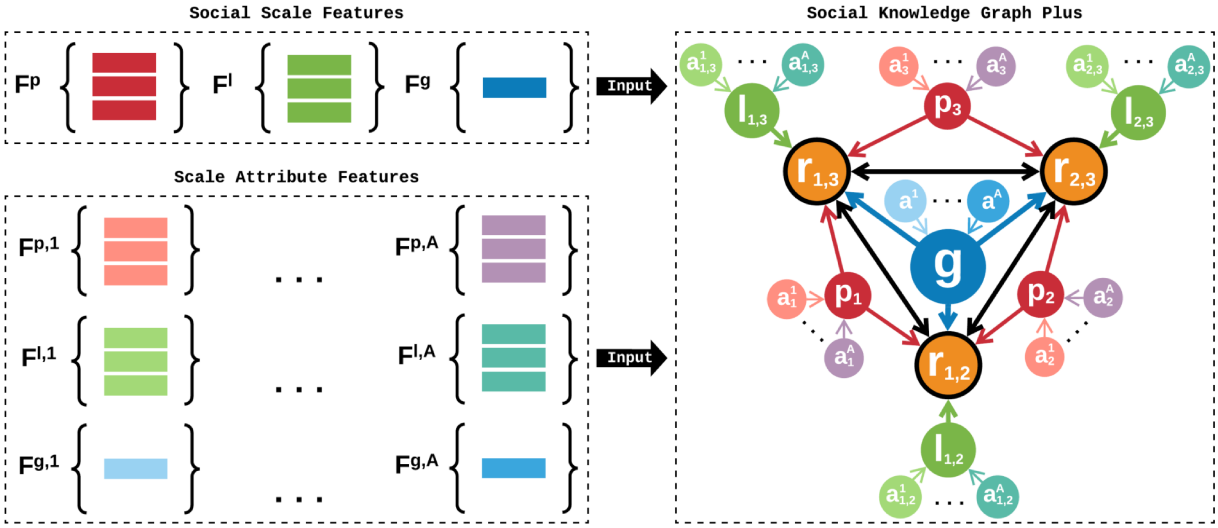


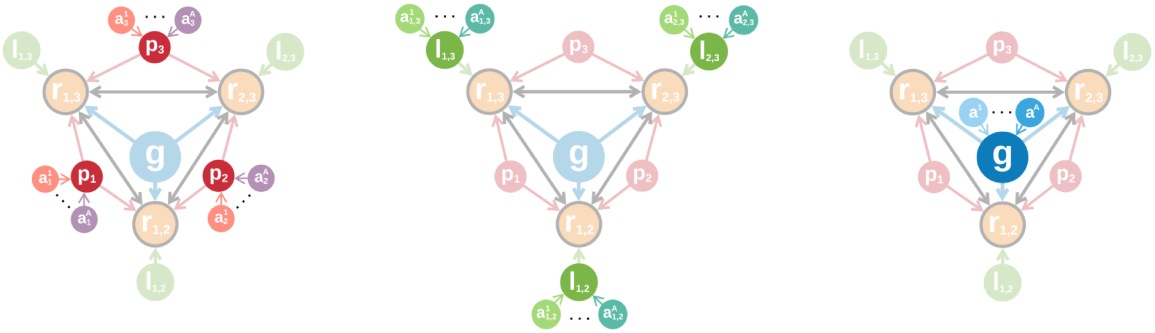
Figure 5.8. Addition of attribute nodes to the SKG constructed in Section 5.2. The initial hidden states of these nodes receive the feature vectors extracted with models pre-trained for the chosen attributes.

Next, an attribute node of the corresponding type is added to the graph for each feature vector extracted. These nodes are linked to their respective social-scale nodes carrying the features from the image patches used to extract attribute information, as shown in Figure 5.8. The edges connecting these nodes also have the same type of their corresponding attribute node, allowing the model to learn specific aggregating strategies for every cue type when combining their information with social-scale features.

More precisely, the overall process of incorporating attributes to the SKG begins by selecting a number $\mathcal{A}^s$ of distinct attribute types tied to a social scale $s = p, l, g$. Next, it is generated a set of attribute feature vectors $F^{s,a} = \{f_i \mid (\forall x_i \in X^s)[f_i = \Phi^a(x_i)]\}$, where $a = 1, 2, ..., \mathcal{A}^s$ is the attribute type extracted from the cropped image patches X^s, using the corresponding pre-trained model Φ^a. The obtained feature vectors are resized by employing attribute-specific FC layers followed by the **ReLU** activation.

After the extraction process, for each node v_i belonging to the social scale s, an attribute node $\mathbf{a}_i^a$ of the cue type a is inserted into the graph, and its hidden state h_i^a is initialized with the corresponding attribute feature vector $f_i \in F^{s,a}$. These attribute nodes are connected to their social-scale nodes by a directed edge $(\mathbf{a}_i^a, v_i)^a$ with the same type, as illustrated by Figure 5.9.

This enhanced version of the graph holds extra information, and for this reason, it is called Social Knowledge Graph Plus (SKG+). It supports the relation reasoning model



(a) Personal **(b)** Local **(c)** Global

Figure 5.9. Step-by-step illustration of the attribute addition process for the SKG+ version. The chosen number of attributes nodes is added and connected to their respective social-scale node by an edge of the same type. Starting from (a) personal attributes, then (b) local attributes, and finally (c) global attributes.

into learning the dependencies between the chosen attributes and also among the features extracted directly from the input image by the SSN. At the end of the relation reasoning process, every scale node will hold a combination of prior and learned information, allowing the use of specific strategies for each social scale.

The possibility of adding multi-scale attributes and constraints to the graph increases the model's descriptive power, helping it learn correct associations between the input data in terms of social relationships, mitigating the generalization problem. Additionally, some benchmarks include relationship labels that are directly associated with specific attributes, such as *mother-child* and *father-child*, which are related to gender and age. In these circumstances, attribute features can offer a huge advantage by allowing direct or indirect inference of the correct classes, as explained in Sections 2.2 and 2.3.

Unlike previous methods, which have a fixed architecture that is strongly attached to the chosen attributes, the proposed graph representation and reasoning model support the insertion of any type or number of traits to every scale, providing a general framework that can replicate the information used in most of the other social relation recognition works. The specific number of chosen attributes $\mathcal{A}$ for each social scale, along with their types and the extraction models employed in this work, are detailed in Section 6.3.

5.4 Social Graph Network

The final module is in charge of performing relation reasoning on the Social Knowledge Graph generated in the previous step, obtaining a classification for each relationship in the input image. This is achieved by a message-passing method that exchanges informa-

tion between neighbor nodes through the graph structure, aggregating their information to generate high-level features inside relation nodes, which are fed to a pair of classifiers, producing the final scores, as shown in Figure 5.10.

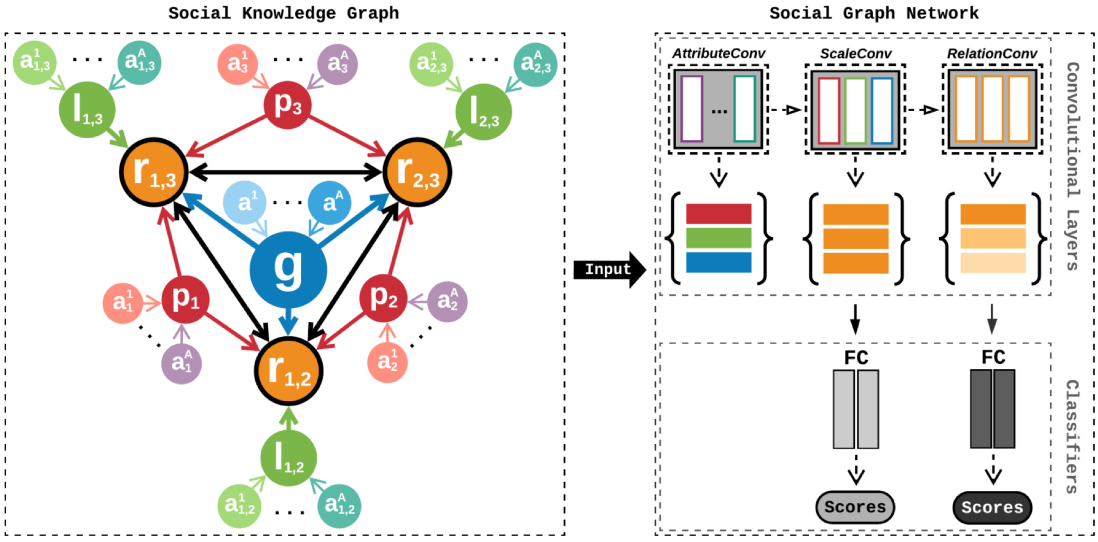
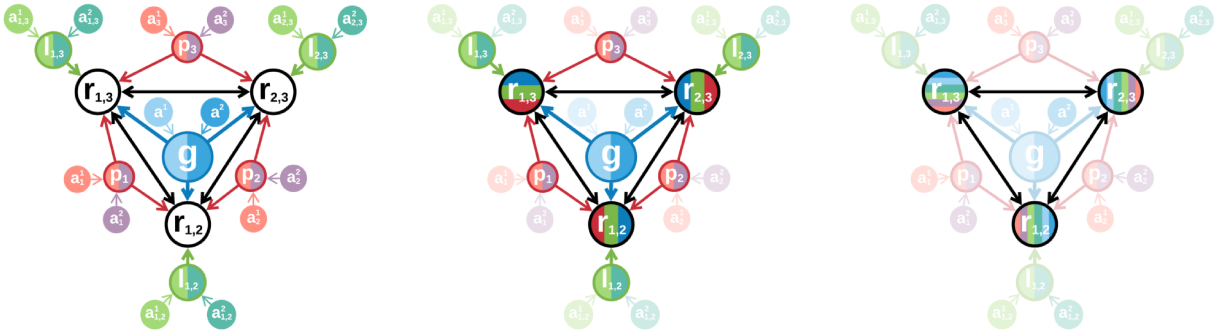


Figure 5.10. A detailed representation of the final module, which is composed of two main stages. The first one performs reasoning on the SKG using the proposed Social Graph Network (SGN), which aggregates all the information carried by the graph, producing high-level features within relation nodes. The second stage receives the generated features, classifying them and outputting the final prediction scores.

Each node type in the graph carries distinct information, which may influence relationship classes in a different way. For example, personal-scale features and attributes like age and gender may be more relevant to identify intimate relationships, such as *couple* and *family*, by capturing strong dependencies that exist between these kinds of relations, as explained in Section 2.3. However, for non-intimate relationships like *commercial* and *professional*, global-scale features and attributes might play a more important role, helping to describe information meaningful for these scenarios, such as working environment objects and group actions.

For this reason, the model needs to learn how to weigh features independently for each type of node, since they can contain scale-specific information, such as social-scale nodes, and also features combining data from multiple sources, like relation nodes. This weighted information is employed in the generation of the message from each node, which is exchanged between their neighbors, and the set of messages received by every node is aggregated and combined with their own information, resulting in a single feature vector that represents the updated hidden state of that node. This information exchange process is known as *message propagation*, and it is guided by the graph structure in the proposed approach.

Each propagation step sends information from the node to its neighbors, which combines these features with their own, allowing the repropagation of the received information from a new source. Every time this process is repeated, the information from the



(a) Attributes **(b)** Social Scales **(c)** Relationships

Figure 5.11. An illustration of the *message propagation* process performed by the SGN, where each color represents a different type of feature. (a) The first step sends attribute features to their respective social-scale nodes. (b) The social-scale features are aggregated in the relation nodes during the second step, which now contains multi-scale and multi-attribute information. (c) Finally, after the third step, all this information is able to reach neighbor relation nodes, capturing multi-scale and multi-attribute interdependencies, and completing the reasoning process. After this procedure, each relation node carries distinct combinations of features for each relationship, represented by different color arrangements.

initial node goes deeper into the graph, as illustrated in Figure 5.11. More specifically, social-scale nodes receive their respective attribute features, which are aggregated and combined with the node's own information extracted by the SSN (Figure 5.11a). After this process, the updated hidden states of the social-scale nodes, which now contain attribute information, are passed as messages to their corresponding relation nodes in the next propagation step (Figure 5.11b).

The information received by these nodes is again aggregated and combined with their previous state, generating a multi-scale and multi-attribute representation for the relationship. A final message passing step (Figure 5.11c) allows the relation nodes to interchange information through relation edges, capturing the dependencies between relationships, as described in Section 2.2. The resulting high-level features are capable of describing each relationship in a distinctive manner, according to their classes.

As mentioned in Section 5.3, the proposed graph structure separates information by its sources, differentiating between nodes and edge types. This approach allows the model to learn specific strategies and proportions to combine this data by employing individual parameters. In other words, the SKG is a heterogeneous graph with at least four different types of nodes and directed edges, and this number increases with each new social-scale attribute added. The Social Graph Network (SGN) is proposed to deal with this variety of information, learning how to weigh and propagate it through the graph, according to its type. This is done by exploiting the rich and intuitive graph structure, applying a specific propagation method for each region, focusing on a particular type of information, and optimizing the reasoning process.

The Social Graph Network consists of three distinct layers, and each one is applied

sequentially to specific types of nodes, producing the information stream depicted in Figure 5.11. Starting with the outermost nodes, which carry attribute data in the SKG+ version, their information has to be passed to the corresponding social-scale nodes. This data is then combined with the current social-scale features extracted from the input image, generating a new hidden state for each scale node. This is done by the *AttributeConv* layer, defined as

$$h_i^s = h_i^s + \mathbf{log}\left(\sum_{a=1}^{\mathcal{A}^s} \mathbf{exp}(\mathbf{tanh}(W^{s,a} h_i^{s,a} + b^{s,a}))\right), \tag{5.4}$$

where $W^{s,a}$ and $b^{s,a}$ are the learned matrix of weights and bias term associated with the attribute type a, $\mathcal{A}^s$ is the total number of attribute types extracted for the social scale s, while h_i^s and $h_i^{s,a}$ are the hidden states of the social-scale node and the corresponding attribute node of type a.

This layer assigns weights for distinct attribute types, allowing the model to learn different combination strategies for each one. The extracted features probably also have different value ranges, depending on the employed models, which can interfere with the learning process. For this reason, the **tanh** activation function is used, fitting the data within the same range. The resulting feature vectors for all attribute types are aggregated using the **LSE** function, providing invariance to node ordering in the graph and normalizing these values, which prevents them from changing the scale features drastically. Finally, the results produced by this process are added to the social-scale features, updating their hidden state with attribute information.

Next, the updated social-scale features have to be combined, generating a multi-scale representation for their respective relationships. Since every relation node $\mathbf{r}_{i,j}$ is connected with two personal nodes $\mathbf{p}_i$ and $\mathbf{p}_j$, one local-scale node $\mathbf{l}_{i,j}$, and to the global-scale node $\mathbf{g}$, their hidden state features, denoted by h_i^p, h_j^p, $h_{i,j}^l$, and h^g, can be directly combined to obtain a representation. This is achieved by the *ScaleConv* layer

$$h_{i,j}^r = \mathbf{ReLU}(W^p(h_i^p + h_j^p) \parallel W^l h_{i,j}^l \parallel W^g h^g), \tag{5.5}$$

where W^p, W^l, and W^g are the learnable weight matrices associated with personal, local, and global-scale features respectively, while bias terms are omitted for brevity. Additionally, the $\parallel$ symbol indicates the concatenation operation and $h_{i,j}^r \in \mathbb{R}^{3\mathcal{H}}$ is the feature vector representing the relationship between the persons i and j. The learnable parameters are shared among features from the same scale, learning to capture the most important aspects of each scope, while personal node features are added, avoiding issues related to the order each person appears.

The third layer is in charge of capturing relationship interdependencies, and therefore it is applied to relation nodes, combining their features with information from their

social neighbors. This task is performed by the *RelationConv* layer

$$h^r_{i,j} = h^r_{i,j} + \mathbf{ReLU}(\mathbf{LayerNorm}(Uh^r_{i,j} + \sum_{\mathbf{r}_{m,n} \in S(\mathbf{r}_{i,j})} Z^{i,j}_{m,n} V h^r_{m,n})), \tag{5.6}$$

where U and V are learnable matrices of weights, and $Z^{i,j}_{m,n}$ is an attention score calculate by

$$Z^{i,j}_{m,n} = \frac{\sigma(\bar{Z}^{i,j}_{m,n})}{\sum_{\mathbf{r}_{m,n} \in S(\mathbf{r}_{i,j})} \sigma(\bar{Z}^{i,j}_{m,n}) + \epsilon}, \tag{5.7}$$

$$\bar{Z}^{i,j}_{m,n} = W h^r_{i,j} + X h^r_{m,n}, \tag{5.8}$$

where W and $X \in \mathbb{R}^{\mathcal{H}}$ are vectors of learnable parameters, ϵ is a small constant for numerical stability and σ is the sigmoid activation function.

For every relationship $\mathbf{r}_{i,j}$, the layer calculates an attention score $Z^{i,j}_{m,n}$ that measures the importance for each one of its neighbor relationships $\mathbf{r}_{m,n}$. This score is used to weigh the features from neighbor relation nodes, which are summed and normalized applying **LayerNorm** [Lei Ba et al., 2016]. The resulting values are added to the hidden state of the current relation node, generating a final representation, which now also includes interdependecies information.

The hidden states of each relation node are fed to a simple classifier composed of two FC layers followed by a softmax function, outputting the final class probability scores for the relation $r_{i,j}$ as

$$\hat{y}_{i,j} = \mathbf{softmax}(W\ \mathbf{ReLU}(Uh^r_{i,j} + b^u) + b^w), \tag{5.9}$$

where W, b^w, U, and b^u represent weight matrices and bias terms. The cost function of an input image $\mathbf{I}$ depicting a set of relationships R with size $\mathcal{R}$ is given by the cross-entropy loss obtained by

$$\mathcal{L} = -\frac{1}{\mathcal{R}} \sum_{r_{i,j} \in R} \sum_{c=1}^{\mathcal{C}} y^c_{i,j} log(\hat{y}^c_{i,j}), \tag{5.10}$$

where $y^c_{i,j}$ is the ground-truth for the relation $r_{i,j}$ with respect to the class c, and $\mathcal{C}$ is the total number of classes.

Finally, an auxiliary classifier with the same structure defined in Equation 5.9 is used to generate class scores from relationship features before the application of the *RelationConv* layer, with the final cost function of the model being the sum of the losses from both classifiers. The purpose of this addition is to inject gradients closer to the backbone models, reducing vanishing gradient issues, similarly to the GoogleNet [Szegedy et al., 2014]. Additionally, the second classifier also helps to reduce overfitting by providing additional regularization to the model, since these intermediary features can be considered

as a pairwise approach. The gradients injected directly into this stage of the model help to assure these features are being learned in the correct direction, since they also have to make sense from a pairwise perspective.

Chapter 6

Experiments and Results

This chapter presents the main benchmark datasets for social relation recognition, along with their evaluation protocols. Additionally, the data processing steps applied to shift the task from a relationship-based to an image-based approach are also described. Next, the baselines and state-of-the-art models selected for comparison are summarized by their fundamental concepts, drawing parallels with the methodology proposed in this work. Finally, details on the implementation process and other relevant choices are also elucidated.

For the experiments used to evaluate the proposed model, the reasoning behind them is explained in their respective sections. The obtained results are presented along with a deep investigation of their quantitative aspects, including an ablation study and the effects of model variations. Finally, a qualitative analysis is conducted, discussing the influence of each design choice and its outcomes by displaying some examples of correct and incorrect classifications.

6.1 Datasets

The proposed approach was evaluated on the two most significant benchmarks for social relation recognition: the (i) PISC, and (ii) PIPA-relation datasets. This section presents a brief explanation of the construction process and statistics for each one, along with a description of their suggested evaluation protocols.

6.1.1 PISC

The People in Social Context (PISC) dataset [Li et al., 2017] is the largest benchmark in the literature focused on social relationships. Based on the relational models theory [Fiske, 1992], it defines hierarchical categories composed of social domains branching into relationships sub-categories, which embed coarse-to-fine aspects of typical social interactions, as illustrated by Figure 6.1.

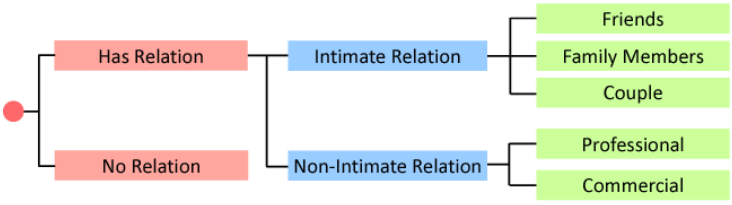
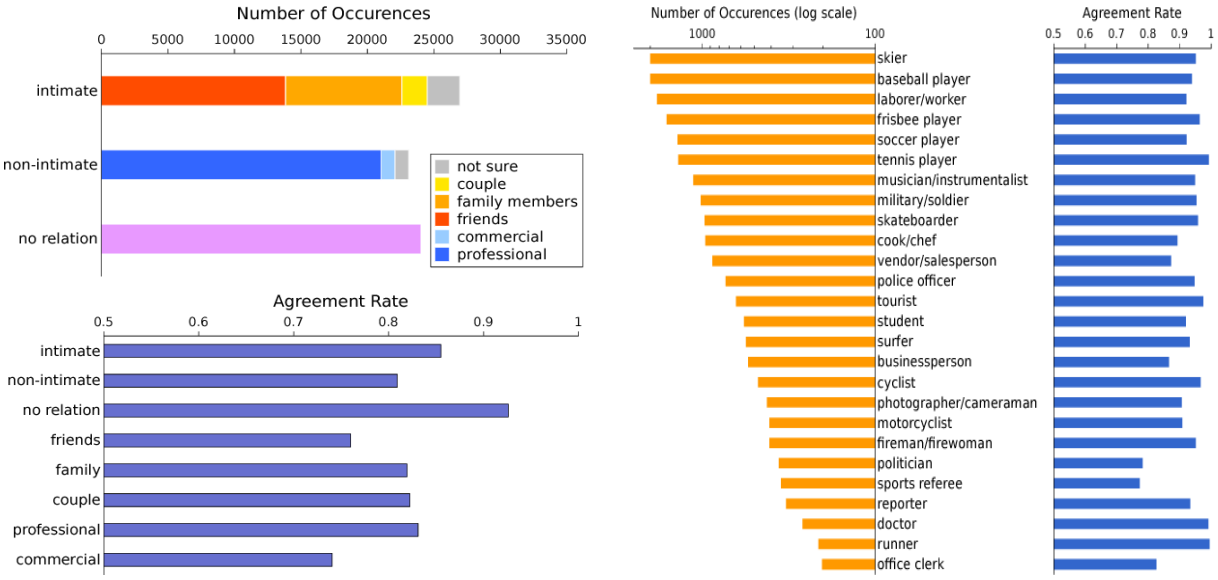


Figure 6.1. Set of hierarchical social relationship categories defined in the People in Social Context (PISC) dataset [Li et al., 2017].

Images from multiple sources were used to build this benchmark, including well-known datasets such as the Visual Genome [Krishna et al., 2017], MS-COCO [Lu et al., 2016] and YFCC100M [Thomee et al., 2016]. Additionally, data from other publicly available platforms like Flick, Instagram, Twitter, and search engines such as Google and Bing were also employed. The collected images were filtered to avoid crowds and single persons, while the remaining samples were manually annotated with body bounding boxes, occupation and social relationships labels for all potential individual pairs. Each image was evaluated by at least five persons, and the final label decision was determined by majority voting. Figure 6.2a shows the agreement rating and the number of occurrences for relationship classes during the annotation process.

The dataset consists of 22,670 images, with an average number of 3.11 persons per picture. For social relations, each pair of individuals in an image is considered as a sample, composing a total of 76,568 valid relationships for the whole dataset. Considering occupation, a total of 10,034 images were annotated with recognizable professional activities. Figure 6.2b shows the 26 most frequent occupation categories and their agreement rate.

Social relation recognition experiments on the PISC dataset are performed using two suggested evaluation protocols. The first one focuses on *social domains*, splitting the dataset into three coarse-level categories: *No Relation*, *Intimate Relation*, and *Non-Intimate Relation*. The train set has 55,400 samples, distributed in 16,828 images, with an average of 3.09 persons per picture. For the test set, there are 3,961 samples distributed in 1,250 images, with an average of 3.06 persons per picture. Finally, the validation set has 1,505 samples, distributed in 500 images, with an average of 2.99 persons per picture,



(a) Social relationships **(b)** Occupations

Figure 6.2. Statistics for the PISC dataset [Li et al., 2017]. (a) The number of occurrences and agreement by class for social relationships annotations. (b) The number of occurrences and agreement by class for occupation annotations.

totalizing 60,866 samples for the entire evaluation protocol.

The second evaluation protocol considers fine-grained *social relationships*, represented by 6 classes: *Friends*, *Family Members*, *Couple*, *Professional*, *Commercial*, and *No Relation*. The train set has 49,017 samples, distributed in 13,142 images, with an average of 3.02 persons per picture. For the test set, there are 15,497 samples distributed in 4,000 images, with an average of 3.11 persons per picture. Finally, the validation set has 14,536 samples, distributed in 4000 images, with an average of 3.07 persons per picture, totalizing 79,050 samples for the entire protocol.

6.1.2 PIPA-relation

The People in Photo Album (PIPA) relation dataset [Sun et al., 2017] is one of the first open benchmarks for social relation recognition. It was inspired by the social domain theory [Bugental, 2000], which provides a robust conceptualization of human social life, encapsulating all aspects of interpersonal interactions while also being concrete and specific enough to support a computational model. The theory partitions social life into domains, providing precise definitions for each one and also indicating related social cues, including appearance and behaviors.

Images from the original PIPA dataset [Zhang et al., 2015] were reused to build

this benchmark. The data was collected from Flickr for person recognition, consisting of 37,107 pictures, including 63,188 samples of 2,356 individuals, covering a wide range of social scenarios. It was annotated with head bounding boxes and identity numbers, allowing the identification of the same individual in distinct social situations, delivering an ideal scenario for relation recognition tasks. The original dataset was extended with relationship annotations, using identity information to select person pairs and establish evaluation protocols. Considering that social relations may be subjective, ambiguous, and in some cases even attached to cultural backgrounds, the annotators were selected from different geographic regions, including four continents.

Each instance from this dataset belongs to one of the five *social domains*: *Attachment*, *Reciprocity*, *Mating*, *Hierarchical Power*, and *Coalitional Groups*, and from them, a list of 16 *social relationships* is derived. The resulting PIPA-relation dataset contains 26,915 samples, with a total of 134,556 annotations and up to three labels per instance, selected by each annotator. The final label for each sample is selected by applying a consistency verification method based on the agreement between annotators, which is shown in Figure 6.3.

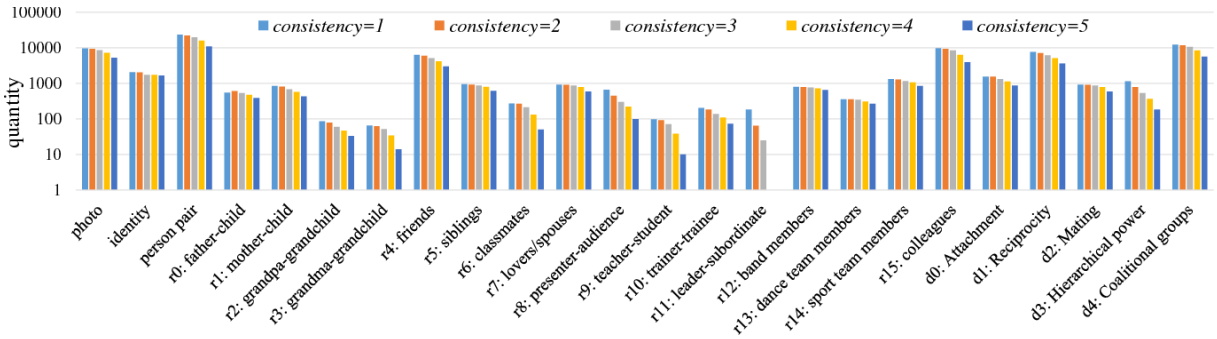


Figure 6.3. Class consistency for *social domain* and *social relationship* classes in the PIPA-relation dataset [Sun et al., 2017].

The experiments on the PIPA-relation dataset are performed using the recommended All-class (AC) protocol, where the training data covers all classes of social relationships and domains. The train set has 13,729 samples, distributed in 5,857 images, with an average of 2.93 persons per picture. For the test set, there are 5,106 samples distributed in 2,452 images, with an average of 2.58 persons per picture. Finally, the validation set has 709 samples, distributed in 261 images, with an average of 2.9 persons per picture, totalizing 19,544 samples for the entire evaluation protocol.

6.1.3 Data Issues

As mentioned in Section 1.3, the social relation recognition task has to work with an ill-posed problem, since not all information required to identify some relations may be presented by the images. This issue becomes explicit in both datasets, where classes like *Friends* and *Family* are impossible to distinguish in some scenarios, opening a margin to multiple interpretations. In this sense, the capabilities of a social relation recognition method are limited to the social theory applied to interpret the relationships, and also to the quality of ground-truth information provided, similarly to other supervised approaches employed on multiple computer vision tasks.

Additionally, modern benchmarks suffer from another recurrent problem on tasks involving machine learning and human beings, which is the diversity of the training data. Most of the images come from homogeneous populations in terms of race, clothing, sexual orientation, and various other physical and cultural traits. This is an obstacle for supervised learning algorithms, resulting in models that are incapable of generalizing to images containing variations in any of these aspects, which often leads to incorrect results.

For these reasons, the objective of this work is not to be assertive about the actual social relationships depicted in the images. Instead, the only purpose is to reproduce the human perception of social relationships reflected in the data, which can be a helpful tool when incorporated into other tasks, as mentioned in Section 1.3.

6.2 State-of-the-art Baselines

This section presents a brief description of some fundamental concepts behind the baselines and state-of-the-art methods used to measure the performance of the proposed framework. A deeper analysis of these approaches was made in Chapter 3, and for this reason, the goal of the following summary is to provide only a quick insight into the progress of the social relation recognition task, helping to confront these methods with their performance in the next sections.

- **Union-CNN** [Lu et al., 2016]: Based on predicate prediction methods, the union region of the individual pair of interest is classified using a CNN model.
- **Pair-CNN** [Li et al., 2017]: This approach employs two identical CNNs with shared weights to extract and classify the features from cropped image patches for both

individuals in the relation.

- **Pair-CNN+BBox+Union** [Li et al., 2017]: Extends the Pair-CNN by adding geometry information from both persons bounding boxes concatenated with Union-CNN features.
- **Pair-CNN+BBox+Global** [Li et al., 2017]: Similar to the previous method; however, the Union-CNN receives as input the whole image instead of the union region containing the pair of interest.
- **Dual-Glance** [Li et al., 2017]: Makes a coarse prediction from the pair of individuals using Pair-CNN+BBox+Union (first glance), incorporating object features combined with an attention mechanism (second glance).
- **Attributes, SVM** [Sun et al., 2017]: Trains multiple convolutional models to identify appearance attributes, which are employed to extract features from each pair, feeding different combinations of these attributes to an SVM classifier.
- **Graph Reasoning Model (GRM)** [Wang et al., 2018b]: Considers a prior knowledge graph for the whole dataset, representing co-occurrences between relationship and object classes, weighted by attention mechanism.
- **Multi-Granularity Reasoning (MGR)** [Zhang et al., 2019]: Combines global information with regional features, employing graphs to represent interactions between individuals and objects.
- **Social Relationship Graph Inference Network (SRG-IN)** [Goel et al., 2019]: Use pre-trained models to extract attributes including age, gender, clothing, activity, and scene features, which are aggregated employing recurrent units.
- **Deep Supervised Feature Selection (DSFS)** [Wang et al., 2020]: Proposes group and dimensional feature selection methods, abstracting individual attribute features obtained by using pre-trained models.
- **Graph Relational Reasoning Network (GR^2N)** [Li et al., 2020]: Employs a graph model for exchanging information between multiple relationships, exploiting their correlations.

6.3 Implementation Details

In this section, essential details about the implementation are provided, allowing the reproduction of the obtained results. Starting from the databases, the procedures required to shift the task from a relation-based to an image-based approach are explained. Next, specifications on the extraction of the social-scale and attribute features used in the experiments are described. Finally, the hyper-parameters and optimization techniques employed are also reported.

6.3.1 Datasets Preparation

No modifications were needed for the PISC dataset [Li et al., 2017], since it was provided with bounding boxes, class labels, metadata annotations, and lists of experiment protocols organized per image. Additionally, the full pictures were also made available, allowing to efficiently crop image-patch regions for the inputs of each social-scale backbone.

On the other hand, the PIPA-relation dataset [Sun et al., 2017] presented evaluation protocols, metadata, and annotations merged into the same file, where each relationship sample is disconnected from the rest of the image, addressing only pair-based labels. Another issue with this dataset is that bounding boxes were also not provided, and instead, only the image regions containing the individuals considered for each sample were given, complicating the generation of the other social-scale image patches.

This issue was circumvented using the source image names and a counter for each person in the same picture provided with the annotations. With this information, it was possible to backtrack head bounding boxes for every person in the PIPA-relation dataset by employing the original PIPA dataset [Zhang et al., 2015] annotations. Finally, after recovering head coordinates, full-body bounding boxes were recalculated in the same way described by Sun et al. [2017]: $3 \times \textit{head width}$ and $6 \times \textit{head height}$. Additionally, the information of persons and labels per image was also obtained from this process, providing all the data necessary for the transition to an image-based approach.

6.3.2 Features Extraction

For a fair comparison, ResNet-101 models [He et al., 2016] pre-trained on the ImageNet [Russakovsky et al., 2015] are employed as backbones for the Social Scale Networks in the first module, since this architecture was also applied for feature extraction in previous works [Li et al., 2017; Wang et al., 2018b; Li et al., 2020]. Each of these backbone models is fine-tuned for their respective social scale by freezing the parameters from the first two residual blocks (*conv1* and *conv2_x*), preserving the weights of these initial layers, since they tend to learn more general features. The last three residual blocks (*conv3_x*, *conv4_x*, and *conv5_x*) are allowed to retrain, adjusting their parameters to learn scale-specific features. Furthermore, the image patches X^p, X^l, and X^g are resized to 224×224 pixels, fitting the input size of the chosen backbone architecture.

Considering that ResNet-101 residual blocks include *batchnorm* layers [Ioffe and Szegedy, 2015], the running mean and standard deviation of the original training are also preserved. This is done because those statistics depend on batch variations, being affected by the number of samples and their values, which can change drastically because of the image-based approach, since some pictures only have one relationship while others can show more than 30. The rapid oscillations in these values invalidate the running mean and standard deviation between batches during the feature learning process, resulting in a non-stationary training, which does not reflect accurate batch statistics through the test phase, degrading the model generalization capabilities. Additionally, this situation can be complicated by the similar inputs for relationships in the same image, as described in Section 2.1. The application of the original statistics helps to mitigate these issues, since social relationship images are in the same domain of the original training data.

The social-scale features are generated by extracting outputs of the last *average pooling* layer of each backbone model, obtaining the feature vectors $F^p \in \mathbb{R}^{\mathcal{P}\times 2048}$, $F^l \in \mathbb{R}^{\mathcal{R}\times 2048}$, and $F^g \in \mathbb{R}^{2048}$, as described in Section 5.2. For the SKG+ version, a value of $\mathcal{A}^p = 3$ is employed, producing the sets of individual appearance attribute features

$$F^{p,1} = \{f_i \in \mathbb{R}^{4096} \mid (\forall x_i \in X^p)[f_i = \mathbf{age}(x_i)]\}, \tag{6.1}$$

$$F^{p,2} = \{f_i \in \mathbb{R}^{4096} \mid (\forall x_i \in X^p)[f_i = \mathbf{gender}(x_i)]\}, \text{ and} \tag{6.2}$$

$$F^{p,3} = \{f_i \in \mathbb{R}^{4096} \mid (\forall x_i \in X^p)[f_i = \mathbf{clothing}(x_i)]\} \tag{6.3}$$

from the *fc7* layer of the **age**, **gender** and **clothing** double-stream Caffe-Net models provided by Sun et al. [2017], which are also used by other works [Goel et al., 2019; Wang et al., 2020], producing fair results.

From local-scale image regions, a number of $\mathcal{A}^l = 2$ attributes are extracted. The first one is obtained from the *global_pool* layer of the Inception-V2 model [Szegedy et al.,

2016] made available by Guo et al. [2020]. The network was pre-trained on the ImageNet [Russakovsky et al., 2015] and fine-tuned on the GroupEmoW database presented in their work, which is focused on group **emotion**. The second attribute is extracted from the *fc7* layer of the CNN-CRF **activity** model [Yatskar et al., 2016] trained on a dataset composed of 126,102 images to recognize 504 activity classes, which was again employed in previous works [Goel et al., 2019; Wang et al., 2020]. The extracted attributes vectors are combined, generating the feature sets

$$F^{l,1} = \{f_{i,j} \in \mathbb{R}^{1024} \mid (\forall x_{i,j} \in X^l)[f_{i,j} = \textbf{emotion}(x_{i,j})]\} \text{ and} \tag{6.4}$$

$$F^{l,2} = \{f_{i,j} \in \mathbb{R}^{1024} \mid (\forall x_{i,j} \in X^l)[f_{i,j} = \textbf{activity}(x_{i,j})]\}. \tag{6.5}$$

Finally, a value of $\mathcal{A}^g = 1$ global-scale attributes are extracted using the set of object bounding boxes provided by Wang et al. [2018b], which includes a number $\mathcal{O}$ of distinct object classes for every image on both datasets. The detected regions are cropped from the whole image patch X^g and fed to a SENet-154 model [Hu et al., 2020], pre-trained on the ImageNet database [Russakovsky et al., 2015] for image classification, generating the set $F^{g,1} \in \mathbb{R}^{\mathcal{O}\times 2048}$ of object features for each patch.

6.3.3 Optimization and Parameters

During the training process, all *linear* layer weights are initialized with Xavier Normal [Glorot and Bengio, 2010] and bias terms are initialized with zeros. The SSN parameters are optimized employing Stochastic Gradient Descent (SGD) with 10^{-4} learning rate, 10^{-4} weight decay, and 0.9 momentum. The SGN parameters are optimized employing AdamW [Loshchilov and Hutter, 2019] with a 2×10^{-4} learning rate and 2×10^{-5} weight decay, and a learning rate decay of 0.98 is applied after each epoch for both optimizers. Finally, the constant ϵ for numerical stability on the *RelationConv* layer is set to 10^{-6}.

Since both employed benchmarks are heavily imbalanced, the weights w_c are applied specifically for each class during the calculation of the cost function. These values can be computed by

$$w_c = \frac{n_t}{n_c \mathcal{C}}, \tag{6.6}$$

where w_c and n_c are the weight and number of samples for the class c, while the total number of samples and classes are represented by n_t and $\mathcal{C}$, respectively.

6.4 Quantitative Results

This section presents the results obtained using the proposed approach, which is compared against state-of-the-art models by employing different metrics. Next, the outcomes are discussed and interpreted, considering design choices and other important aspects of each method.

Following previous works, per-class recall and mean average precision (mAP) metrics are computed for PISC, while accuracy is employed for the PIPA-relation dataset. As shown in Tables 6.1, 6.2 and 6.3, the proposed model was capable of surpassing all previous methods, achieving a new state-of-the-art on both benchmarks.

Table 6.1. Comparison of recall-per-class and mean average precision (mAP) metrics for the proposed methodology against the state-of-the-art on the *relationship* split of the PISC dataset.

Methods	**Relationship**						
	Friends	Family	Couple	Professional	Commercial	No Rel.	**mAP**
Union-CNN [Lu et al., 2016]	29.9	58.5	70.7	55.4	43.0	19.6	43.5
Pair-CNN [Li et al., 2017]	30.2	59.1	69.4	57.5	41.9	34.2	48.2
Pair-CNN+BBox+Union [Li et al., 2017]	32.5	62.1	73.9	61.4	46.0	52.1	56.9
Pair-CNN+BBox+Global [Li et al., 2017]	32.2	61.7	72.6	60.8	44.3	51.0	54.6
Dual-Glance [Li et al., 2017]	35.4	68.1	76.3	70.3	57.6	60.9	63.2
GRM [Wang et al., 2018b]	59.6	64.4	58.6	76.6	39.5	67.7	68.7
MGR [Zhang et al., 2019]	**64.6**	67.8	60.5	76.8	34.7	70.4	70.0
SRG-IN [Goel et al., 2019]	-	-	-	-	-	-	71.6
GR^{2}N [Li et al., 2020]	60.8	65.9	**84.8**	73.0	51.7	70.4	72.7
SGN(SKG)	57.0	**75.8**	62.1	**83.9**	48.3	59.8	75.2
SGN(SKG+)	49.4	70.5	74.6	76.5	**59.6**	**74.6**	**75.2**

Table 6.2. Comparison of recall-per-class and mean average precision (mAP) metrics for the proposed methodology against the state-of-the-art on the *domain* split of the PISC dataset.

Methods	**Domain**			
	Intimate	Non-Intimate	No Rel.	**mAP**
Union-CNN [Lu et al., 2016]	72.1	81.8	19.2	58.4
Pair-CNN [Li et al., 2017]	70.3	80.5	38.8	65.1
Pair-CNN+BBox+Union [Li et al., 2017]	71.1	81.2	57.9	72.2
Pair-CNN+BBox+Global [Li et al., 2017]	70.5	80.0	53.7	70.5
Dual-Glance [Li et al., 2017]	73.1	**84.2**	59.6	79.7
GRM [Wang et al., 2018b]	81.7	73.4	65.5	82.8
GR^{2}N [Li et al., 2020]	81.6	74.3	70.8	83.1
SGN(SKG)	**88.3**	67.6	69.7	85.6
SGN(SKG+)	83.6	69.8	**75.6**	**85.8**

For the PISC dataset, the proposed method obtained an mAP of 75.23% for *relationship* classes and 85.78% for the *domain* split, with an increase of 2.53 and 2.68

percentage points against the highest previous scores. The model achieves a better performance improvement on the three classes split, probably because of the global-scale features adding context information to the relationships, which is not considered by the previous state-of-the-art method [Li et al., 2020], since it employs only personal-scale image patches. For the six-relationships split, the proposed method shows a similar improvement, suggesting that it is also more efficient in capturing the fine-grained information needed to differentiate between more classes. These results may be associated with relative information provided by local-scale features, which is also neglected by the previous state-of-the-art.

Table 6.3. Comparison of accuracy metric for the proposed methodology against the state-of-the-art on the PIPA-relation dataset.

Methods	**Accuracy (%)**	
	Domain	Relationship
All Attributes, SVM [Sun et al., 2017]	67.8	57.2
Dual-Glance [Li et al., 2017]	-	59.6
DSFS-Dimensional [Wang et al., 2020]	-	61.5
GRM [Wang et al., 2018b]	-	62.3
MGR [Zhang et al., 2019]	-	64.4
GR^2N [Li et al., 2020]	72.3	64.3
SGN(SKG)	73.0	66.7
SGN(SKG+)	**73.8**	**68.0**

For the PIPA-relation dataset, the proposed methodology also provides an improvement over previous models on both evaluation splits, showing an increase of 3.6 and 1.4 percentage points against Li et al. [2020] and Zhang et al. [2019], respectively. The *relationship* protocol covers 16 classes, and this is where the model produced the most significant improvement, suggesting that previous methods lack the high-level information necessary to recognize more specific relationships, which can be achieved by the proposed method. It is also noticeable how similar are the improvements values over previous works for both splits of each dataset, showing the consistency of the methodology.

Additionally, the use of the same backbone and attribute extraction models employed in previous works [Sun et al., 2017; Li et al., 2017; Wang et al., 2018b; Zhang et al., 2019; Goel et al., 2019; Wang et al., 2020; Li et al., 2020] provided a fair comparison, which indicates that the proposed framework was more efficient in capturing dependencies between the training data, resulting in a higher generalization capacity. The main reasons behind these numbers are probably the combination of all social scales and the preservation of the social structure, allowing the extraction of information from other relationships, which is further investigated in the following sections.

6.4.1 Model Variations

In this section, the consequences of model variations are estimated on both datasets. Initially, multiple hidden state dimension values ($\mathcal{H}$) for the SGN module are evaluated and discussed, followed by an investigation on the outcomes produced by different attribute aggregation methods.

6.4.1.1 Hidden State Dimension

In this experiment, the influence of the hidden state size is evaluated. As stated in Section 5.4, the features obtained from social-scales and attributes extraction models are required to have identical dimensions for the graph network to process them. This is accomplished by downscaling these vectors to an equivalent size, which is defined as the hidden state dimension of the node features.

For this reason, the value of $\mathcal{H}$ is directly associated with the capacity of the proposed model. Small hidden sizes may not be enough to represent the learned features in a distinctive way, while large feature dimensions will result in an excessive number of learnable parameters, slowing down the training process and increasing the chance of overfitting. The obtained results for multiple variations on the value of $\mathcal{H}$ are shown in Table 6.4.

The ideal value for $\mathcal{H}$ is between 256 and 512 on both datasets, while increasing or decreasing beyond this interval started to degrade the performance. Since a hidden state size of 512 performed slightly better in general, it was chosen as the hidden state dimension of the model for the following experiments.

6.4.1.2 Attributes Aggregation Function

This experiment consists of replacing the attribute aggregation method in the *AttributeConv* layer for the SKG+ with other functions, measuring their outcomes. The purpose of this layer is to combine the multiple sets of features received from various sources, producing a single vector that is used to update the representation of the corresponding scale node, as explained in Section 5.4.

Table 6.4. The effects of different hidden state dimension ($\mathcal{H}$) values on the model's performance, estimated by accuracy and mean average precision (mAP) metrics on both datasets.

$\mathcal{H}$	**PISC**			
	Accuracy (%)		**mAP (%)**	
	Domain	Relationship	Domain	Relationship
64	75.34	65.59	84.33	73.80
128	75.67	65.44	84.65	74.30
256	76.01	**68.39**	85.01	74.79
512	**76.02**	66.32	**85.56**	**75.21**
1024	75.08	67.76	84.33	74.84
$\mathcal{H}$	**PIPA-relation**			
	Accuracy (%)		**mAP (%)**	
	Domain	Relationship	Domain	Relationship
64	69.49	61.73	54.33	30.84
128	71.92	62.42	55.10	27.96
256	71.54	64.08	**55.64**	**33.79**
512	**72.95**	**66.67**	55.03	33.13
1024	70.45	60.36	52.85	30.70

In this sense, a different feature representation may be generated depending on the employed function, influencing the final performance of the model. The obtained results are shown in Table 6.5, including the original **LSE** function, along with **mean** and and **sum** aggregation methods.

Table 6.5. The effects of different attribute aggregation methods on the model's performance, estimated by accuracy and mean average precision (mAP) metrics on both datasets.

Aggregation	**PISC**			
	Accuracy (%)		**mAP (%)**	
	Domain	Relationship	Domain	Relationship
mean	75.91	68.01	85.00	75.08
sum	76.71	**68.59**	85.59	75.20
LSE	**76.81**	67.91	**85.78**	**75.23**
Aggregation	**PIPA-relation**			
	Accuracy (%)		**mAP (%)**	
	Domain	Relationship	Domain	Relationship
mean	71.58	66.41	56.05	**35.74**
sum	71.41	66.00	58.43	34.82
LSE	**73.78**	**68.00**	**58.80**	34.18

The **mean** function evens input features, generating more stable results. However,

sometimes it can reduce the importance of high feature values, resulting in a loss of distinctive power for strong activations. Similarly, the **LSE** function also evens out the set of features, but it is more sensitive to variations, which may help preserve critical information. On the other hand, the **sum** function captures an overall description of the input features, fully incorporating all feature values into the final representation, which can be a good option in some circumstances, preserving important activations.

Overall, the best results were obtained by the **LSE** function, suggesting that it provides an adequate balance between the preservation of distinctive features and the detection of important patterns from the input data. In second place, the **sum** function generated close results, probably due to its feature distinction capabilities, which provided slightly better metrics on PISC. Additionally, this same factor may also explain the poor performance of the **mean** function on both datasets.

6.4.2 Ablation Study

This section presents an analysis of the contribution provided by particular modules and concepts behind the proposed methodology. The conducted experiments estimate the significance of each social scale, multiple graph versions, and different relationship connections for the model's performance. Additionally, a deep discussion is also conducted to interpret the obtained results.

6.4.2.1 Social Scales

In this experiment, the importance of each social scale is investigated by generating different versions of the Social Knowledge Graph for every possible combination between them. The resulting graphs only contain nodes and features associated with the specified scales, allowing the evaluation of their impact on the final metrics for both datasets, which are shown in Table 6.6.

Separately, the personal scale produced the best outcomes in all scenarios, suggesting that individual appearance features offer the most distinguishable traits for social relation recognition. In second place, the local scale also generated similar numbers, which were probably negatively affected by destructive interference from nearby persons. Finally, the global scale reported the worse results, due to the fact that a single image

Table 6.6. The effects of each social scale on the model's performance, estimated by accuracy and mean average precision (mAP) metrics on both datasets.

Social Scales	PISC			
	Accuracy (%)		mAP (%)	
	Domain	Relationship	Domain	Relationship
Personal	73.27	64.60	81.87	70.69
Local	72.92	63.85	81.23	70.34
Global	66.42	56.20	73.09	62.31
Personal + Local	75.07	66.30	83.52	73.74
Personal + Global	75.65	66.20	84.47	73.05
Local + Global	73.58	65.41	82.34	71.53
Personal + Local + Global	**76.02**	**66.32**	**85.56**	**75.21**
Social Scales	**PIPA-relation**			
	Accuracy (%)		mAP (%)	
	Domain	Relationship	Domain	Relationship
Personal	70.19	60.34	52.90	30.18
Local	68.43	58.30	51.36	28.08
Global	68.12	54.56	47.48	24.50
Personal + Local	71.52	64.36	55.83	32.21
Personal + Global	71.29	64.24	54.82	30.58
Local + Global	71.41	62.81	53.00	33.47
Personal + Local + Global	**72.95**	**66.67**	**55.03**	**33.13**

can contain different classes of relationships, causing severe inconsistencies during the learning process if only these types of features are considered.

All the obtained results are in line with the hypothesis presented in Section 2.1, and also with the predictions of the domain-based theory [Bugental, 2000], in respect to the importance of individual appearance features for the identification of social relations. Additionally, the outcomes are also coherent from a logical point of view, since relationships are primarily defined by persons and their actions, while the environment can only offer additional information.

The results for different scale combinations indicate some level of information overlapping between all of them, which can be aggravated in images where individuals are close to each other, or in scenes with no additional context information. Local and global-scale patches appear to carry the highest overlap, which can be more prevalent in close distance images of a small number of persons, where the regions employed in these two scales can become very similar. Naturally, personal and global-scale patches seem to have the lowest overlap, since they are the most distant scopes.

Nonetheless, the experiment revealed a significant performance increase produced by the use of the three scales in comparison with other versions of the model employing

only one or several combinations of two scales. The obtained results endorse the hypothesis presented in Section 2.1, suggesting that each different scope provides the model with meaningful, and most importantly, complementary information, which can be accessed only from specific social-scales perspectives.

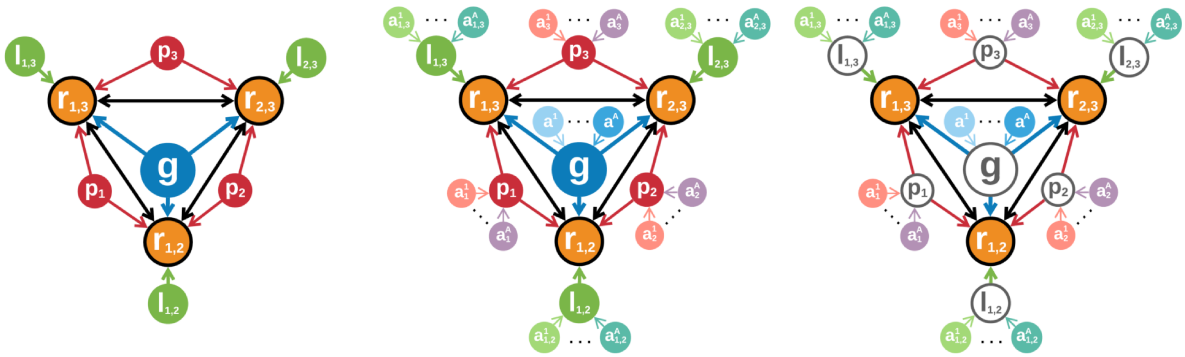
6.4.2.2 Graph Versions

The intent behind this experiment is to evaluate the contribution of the attribute features. This is done by comparing three versions of the Social Knowledge Graph, which are illustrated in Figure 6.4, and constructed in the following ways:

SKG This is the default version of the graph built with all the social-scale nodes and edges, as described in Section 5.3. The results obtained by the SKG will serve as a baseline for the other types of graphs in this experiment.

SKG+ In this version, the original SKG is extended with specific attribute nodes connected by edges of the same type to their corresponding social-scale nodes. This version allows to measure the contribution provided by attribute features to the final model's performance. The graph generation process is presented in Section 5.3.

SKG- This graph variation contains only attribute features, allowing their evaluation separately from the social-scale features, and providing insight into their descriptive capacity for the task. This version is built by removing the social-scale information from the SKG+, and replacing the hidden states of these nodes with zeros vectors, forcing the model to learn based on the attribute nodes alone.



(a) SKG **(b)** SKG+ **(c)** SKG-

Figure 6.4. A representation for each version of graph tested in this experiment. (a) The default SKG. (b) The SKG+ version containing attribute nodes. (c) The SKG- version, removing social-scale information.

As shown in Table 6.7, the addition of attribute information to the SKG improved the performance of the model on both datasets, especially the accuracy for the PIPA-relation dataset. On the other hand, the SKG- underperformed in all scenarios, particularly for the PISC dataset, showing precision values almost 3 percentage points lower than the default version of the graph.

Table 6.7. The effects of different graph versions on the model's performance, estimated by accuracy and mean average precision (mAP) metrics on both datasets.

Graph	**PISC**			
	Accuracy (%)		**mAP (%)**	
	Domain	Relationship	Domain	Relationship
SKG-	74.70	66.78	83.74	72.81
SKG	76.02	66.32	85.56	75.21
SKG+	**76.81**	**67.91**	**85.78**	**75.23**
Graph	**PIPA-relation**			
	Accuracy (%)		**mAP (%)**	
	Domain	Relationship	Domain	Relationship
SKG-	72.11	66.08	58.25	33.59
SKG	72.95	66.67	55.03	33.13
SKG+	**73.78**	**68.00**	**58.80**	**34.18**

The benefits of adding attribute information are probably connected with the methodology applied in the construction of the dataset. For example, the PISC dataset does not have labels directly related to attributes (e.g., *friends*, *family*, *commercial*), and in this case, the model has to learn secondary dependencies between classes and attributes. On the other hand, the PIPA-relation dataset has classes like *father-child*, *mother-child*, *grandfather-grandchild*, and *grandmother-grandchild*, which can be directly associated with age and gender attribute values. These factors may explain the reason why the performance increase was higher for the PIPA-relation dataset.

Additionally, the results were probably also influenced by the sources of the attribute features, since some of them were obtained using the models provided by Sun et al. [2017], which were trained with age and gender labels annotated on the PIPA dataset by Oh et al. [2020]. Although these models may display satisfactory metrics on the original dataset, their generalization capability is deficient due to the restricted number of labels provided. This explains the more significant improvements and also the reason why the SKG- had precision numbers even higher than the default version for the PIPA-relation dataset while performing so poorly on PISC.

Either way, this experiment indicates that adequate attribute features can help to improve the performance, suggesting that they support the model into learning associations between specific aspects and relationships classes, or even acting as constraints in

particular situations, as mentioned in Section 2.3. However, inadequate features can decrease the performance, adding noise to the model and hindering the learning process. The best option would be employing attribute extraction models trained in the same data as the social-scale feature models, but this is not possible since there are no such annotations for PISC, and the small set of attribute labels for the PIPA dataset is insufficient.

6.4.2.3 Relationships Connections

This final ablative experiment evaluates the mechanism proposed to capture interdependencies between relationships in the same image, the *social neighbor* edges. The test is conducted by applying minor modifications to the default SKG, while observing their effects on the model's performance.

Some social relationships can be directly correlated, and this information is exploited in the SKG by inserting edges between relation nodes. The significance of this approach is evaluated by modifying these edges in the following ways:

Social neighbors This is the default version of the graph implementing *social neighbors* connections, as described in Section 5.3. The results provided by this version are used as baselines for the other edge variations in this experiment.

Full neighbors In this version, relation edges are added between all the relationship nodes within the same image. As explained in Section 5.3, some pictures do not have labels for all possible relationships, and this variation will show the effects of considering all the available unlabeled information.

No neighbors All relation edges are removed from the default graph for this modification, isolating relationship nodes from each other, since this is the only path of communication between them. This experiment allows the determination of how significant are the interdependencies between social relations.

The results in Table 6.8 indicate that *social neighbors* connections worked as intended, providing the best outcomes from all other options. In a close second place, the *full neighbors* variation suggests that the number of missing annotations in the PISC dataset is not detrimental to the performance, even adding some accuracy. However, the performance increments were significant for the PIPA-relation dataset, which contains most of the unlabeled samples.

In general, the experiment demonstrates that relationship interdependencies can be exploited to improve relation recognition models, as mentioned in Section 2.2. However,

Table 6.8. The effects of different relationship connections on the model's performance, estimated by accuracy and mean average precision (mAP) metrics on both datasets.

Edges	PISC			
	Accuracy (%)		mAP (%)	
	Domain	Relationship	Domain	Relationship
No neighbors	74.11	65.34	82.79	71.88
Full neighbors	**76.89**	**66.78**	85.45	74.89
Social neighbors	76.02	66.32	**85.56**	**75.21**
Edges	**PIPA-relation**			
	Accuracy (%)		mAP (%)	
	Domain	Relationship	Domain	Relationship
No neighbors	71.52	61.99	52.95	31.70
Full neighbors	72.64	64.92	54.22	**33.19**
Social neighbors	**72.95**	**66.67**	**55.03**	33.13

the information has to be directed to the correct relation nodes; otherwise, it can add noise to the classification, as explained in Section 5.3.

The *no neighbors* version is equivalent to previous methods [Sun et al., 2017; Li et al., 2017; Wang et al., 2018b; Zhang et al., 2019; Goel et al., 2019; Wang et al., 2020], which employed pair-based approaches, completely isolating relationships from each other. The results obtained by this variation are close to the numbers presented by those works, indicating the consistency of the proposed framework. Since all of these methods use only information from two social scales, the percentage gains over them can be explained by the addition of the third scale, also highlighting the efficiency of the methodology even for a pair-based paradigm.

6.5 Qualitative Results

In this final section, fundamental concepts behind the proposed model are investigated using classification samples from the PISC and PIPA-relation test sets. These examples are produced by some of the models generated for the experiments from the previous section. The purpose of the conducted analysis is to associate design choices with the achieved results, providing insight into the model's decision-making process.

For comparison, pairs of correct and incorrect classifications are presented, along with discussions about the mechanisms behind these outcomes. Initially, images generated

with model interpretability techniques are shown, offering visualizations of the information extracted by each social-scale backbone model. Finally, samples of the influence caused by relationship interdependencies and attribute features are examined by confronting the outputs generated with different modules and graph versions.

6.5.1 Social Scale Features

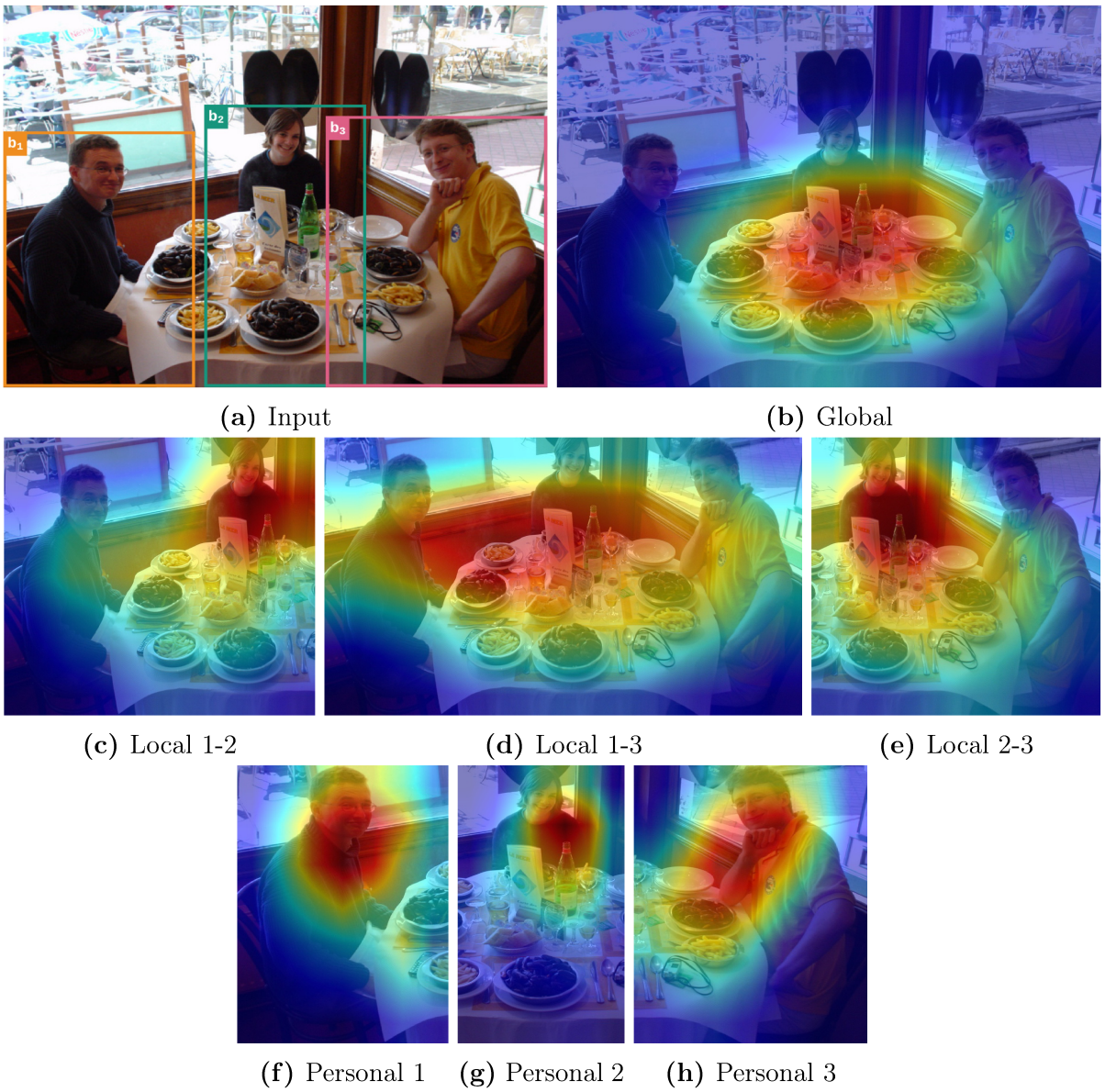
In this section, the social-scale backbones are investigated in an attempt to understand the features they learn. This is done by employing the Gradient-weighted Class Activation Mapping (GradCAM) [Selvaraju et al., 2017], a well-known model interpretability technique that can help to decide if these networks behave as intended, selecting information from image regions that are associated with their respective social scales.

The GradCAM method consists of tracing gradient values back to the final activation maps generated by convolutional models and using them to weigh these activations, which generates a heatmap that can be projected into the input image. The resulting visualization highlights the regions that produced the strongest activations, providing insight into the features learned by the model and allowing the identification of particular cues that may be relevant to the model's decision-making process.

For the correctly classified samples, the model worked precisely as described in Section 2.1, extracting individual traits from personal-scale patches, relative pairwise features from local-scale regions, and general features from the global-scale image. The first example contains three relationships between three persons, the input image with the given bounding boxes, and the GradCAM images extracted from each social-scale model, which are shown in Figure 6.5.

Starting with the global scale, it is noticeable that the strongest activations come from the table, suggesting that the model associated with this scale focuses on general context features, such as objects and background, while the regions containing persons were ignored in this sample. For local-scale patches, the corresponding model covered image regions connecting the bodies of the involved persons, which is especially clear in Figure 6.5d, where the activations reached the individuals on both ends of the table. Finally, as seen from the last three images, the personal-scale backbone extracted individual features from face and torso regions.

All relationships from this sample were correctly classified by the final model as *friends*. The fact that each social-scale backbone focuses on distinct regions and learns different features suggests that they work as expected, with each one offering complementary information. For this sample, the global and local-scale features seemed to be the



(a) Input **(b)** Global

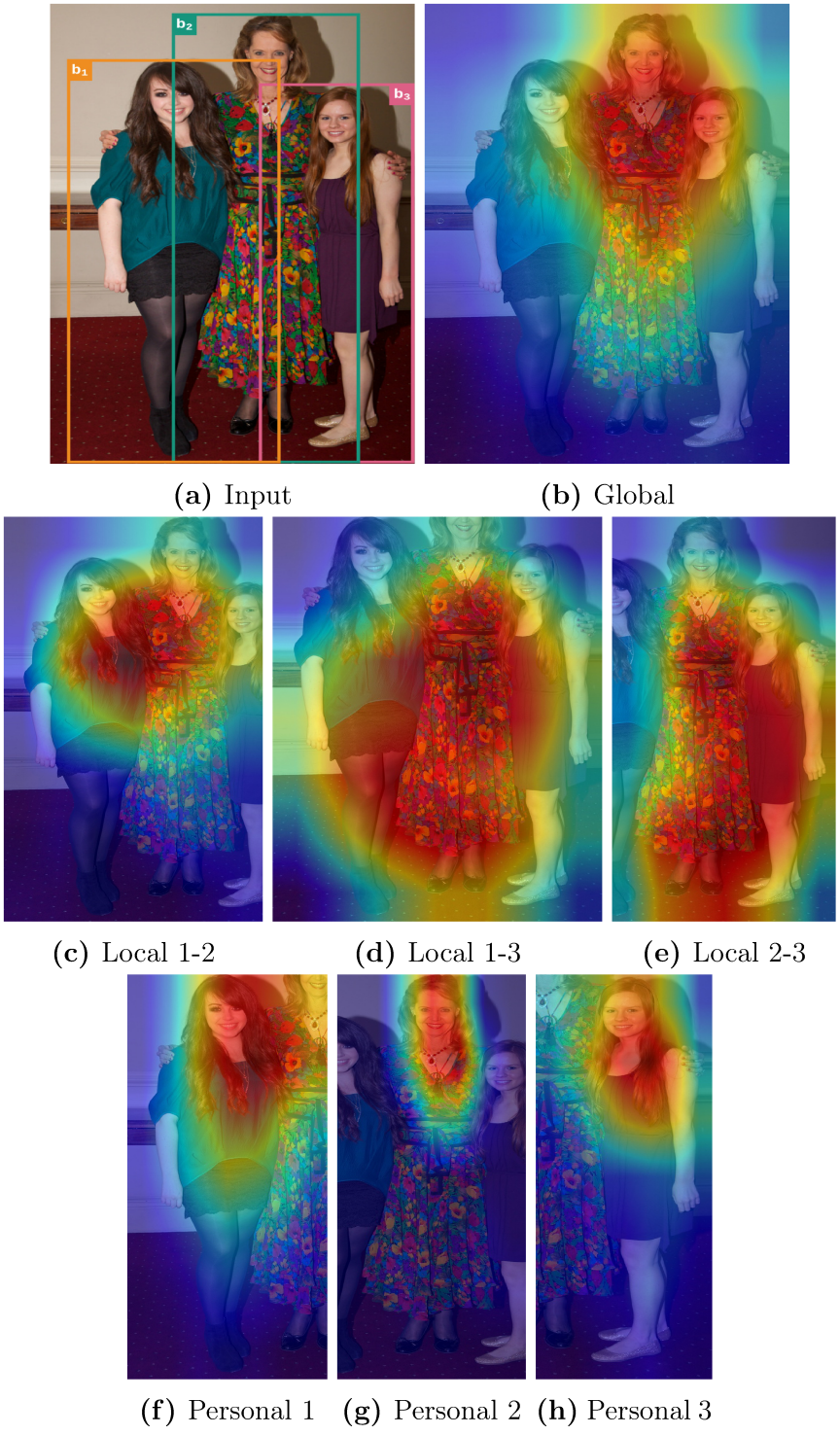
(c) Local 1-2 **(d)** Local 1-3 **(e)** Local 2-3

(f) Personal 1 **(g)** Personal 2 **(h)** Personal 3

Figure 6.5. Example showing the image regions that produced the strongest activations for each social-scale backbone. (a) Input image and the given bounding boxes. (b) The global-scale model focused on the table and other objects. (c)(d)(e) The local-scale backbone considered relative information, extracting features from the image regions covering each pair of persons. (f)(g)(h) The personal-scale network produced individual features focused on face and torso regions.

most significant, capturing the contextual information from the image. However, when the picture contains only persons, and context is not relevant to identify the relationships, it is necessary to employ personal and relative traits to achieve the correct classification. The importance of these two scales and the role of each one is better demonstrated in the second example, shown in Figure 6.6.

In this image, the local-scale model correctly determines the pairs of individuals for each relationship, as it can be seen from Figures 6.6c, 6.6d, and 6.6e, where the strongest activations covers the regions containing the bodies of both persons involved,



(a) Input **(b)** Global

(c) Local 1-2 **(d)** Local 1-3 **(e)** Local 2-3

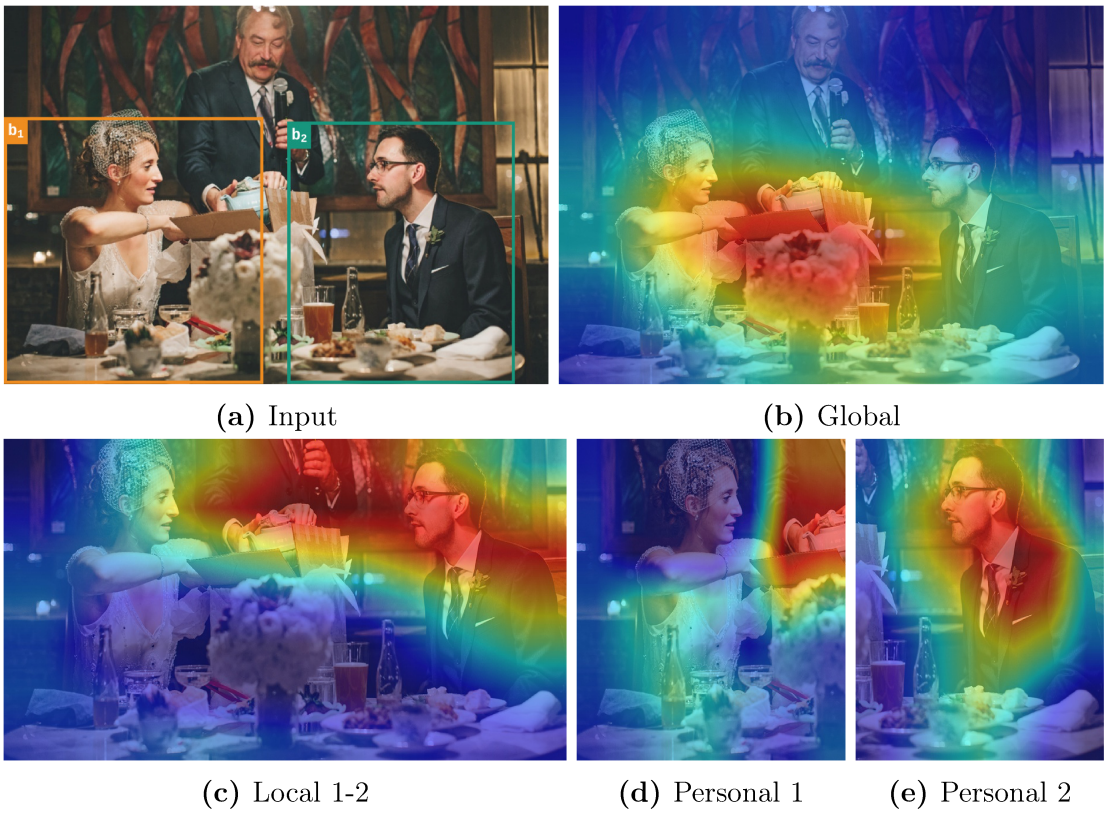
(f) Personal 1 **(g)** Personal 2 **(h)** Personal 3

Figure 6.6. Example where the model was able to correctly determine each relationship pair from (c)(d)(e) local-scale images and every individual from (f)(g)(h) personal-scale patches. This was possible due to the synergy between these two scales, which produced the correct classifications of *mother-child* for 1-2, *siblings* for 1-3, and *mother-child* for 2-3.

focusing on clothing features. The correct localization of these pairs was possible because of the interaction with personal-scale features, which again worked as intended, extracting individual traits, as shown in Figures 6.6f, 6.6g, and 6.6h.

All the relations in this example were properly classified as *mother-child* for 1-2,

siblings for 1-3, and *mother-child* for 2-3, showing that the synergy between these scales is essential for social relationship recognition tasks, and none of the previous works is able to gather information from all of these sources. The missing information becomes even more significant when the model relies only on local-scale features, resulting in erroneous classifications caused by the noise added after fusing personal image regions. In both examples presented, the local-scale image for the most distant individuals also included the person in the middle. This interference problem is described in Section 2.1, and it can be highly detrimental to the model. However, this was not the case for the presented examples because the correct individual features were also considered, and the relationships classes were correlated, resulting in a non-destructive interference. A counter-example containing destructive interference is shown in Figure 6.7.



(a) Input **(b)** Global

(c) Local 1-2 **(d)** Personal 1 **(e)** Personal 2

Figure 6.7. Example of destructive interference exerted by the ceremonialist over the *lovers/spouses* relationship, as shown by (c) local and (d) personal-scale activations. In this case, the relationship was mistakenly classified as *friends*.

In this sample, only the relationship for the couple is considered, and it was incorrectly classified as *friends*. As it can be seen in Figures 6.7c and 6.7d, the interferences caused by the ceremonialist in the local and personal-scale image patches were detrimental. The model was induced to associate the image of the husband with the ceremonialist instead of the wife, producing an outcome that, although reasonable for the regions actually considered, is incorrect for the target relationship. Additionally, the global-scale backbone worked as intended, focusing on the gift box and other contextual data. How-

ever, for this sample, the information required to achieve an appropriated classification can only come from personal and local-scale features, defining the meaning of the global-scale information.

Finally, Figure 6.8 shows an example employing the models generated for the ablation study in Section 6.4, demonstrating how the output can change according to the information considered by each social-scale model. As expected, the personal-scale backbone correctly classifies the *father-child* relationship, and the local scale produces a close *grandfather-grandchild* result, only mistaking the father's age. Individual traits are overlooked by the global-scale model, but it also outputs a family-related classification of *siblings*. This outcome probably occurred due to the features learned from other similar images taken by the same groups of people that are included on the PIPA-relation dataset. The final model classification, combing all three scales, was also correct, suggesting that the *father-child* features from the personal scale were the most relevant.

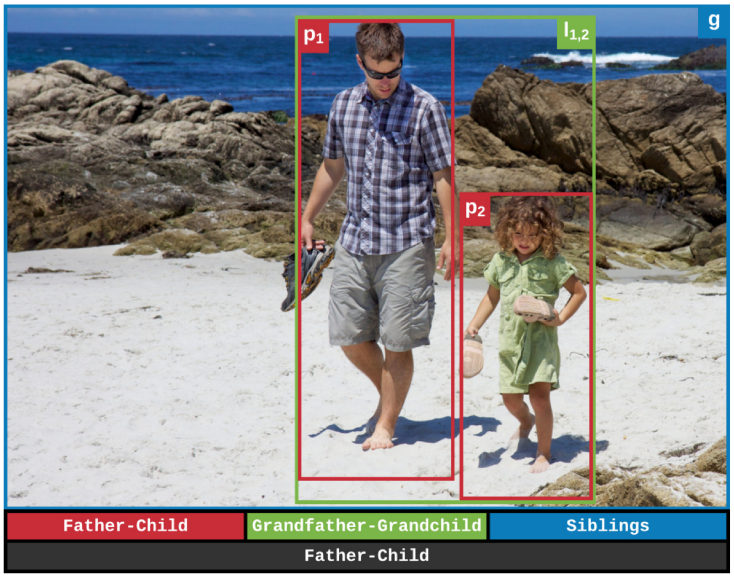


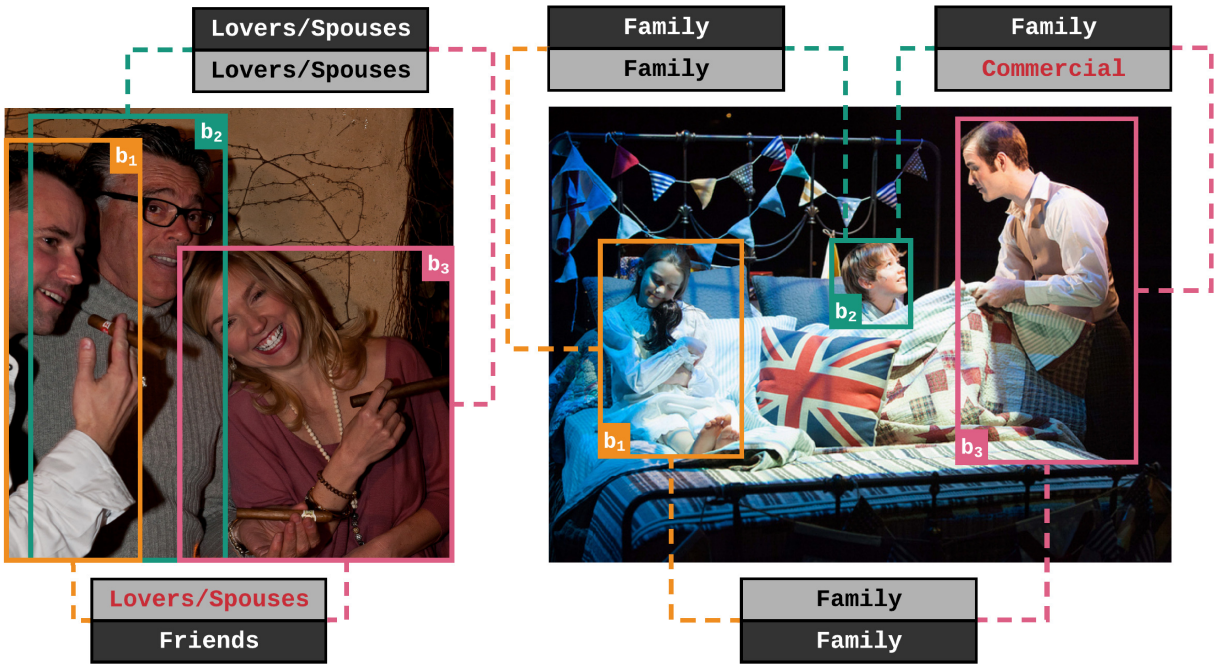
Figure 6.8. Example of how the results can change according to the information captured by each social scale. The image was classified as *siblings*, *grandfather-grandchild*, and *father-child* using global, local, and personal-scale features, respectively. The final model was able to preserve the correct *father-child* output after combining the data from these three sources.

During the conduction of the experiments, it was possible to notice that many of the incorrect classifications made by the model were caused by destructive interference, especially for the PIPA-relation dataset, where bounding boxes are obtained using head proportions, as described in Section 6.3. This method works for standing up poses, but it produces much larger bounding boxes in other situations, adding noise to the image patch, as shown in some of the examples. Some of these interferences could be avoided by adjusting bounding boxes coordinates, and this modification alone could possibly improve even further the model's performance.

6.5.2 Relationships Interdependencies

This section presents examples where the effects of correlation between relationships classes can be exploited, as described in Section 2.2. The obtained results indicate that the proposed model is capable of capturing interdependencies by comparing classifications generated with the scores produced by the auxiliary and final classifiers. Additionally, the implementation of the *full neighbors* model from Section 6.4 is also employed in this experiment, offering further insight into the influence of *social neighbor* connections.

As stated in Section 5.4, the auxiliary classifier receives features generated by the *ScaleConv*, which considers social-scale data from their corresponding relationships, capturing only intradependencies. Next, these feature vectors are fed to a *RelationConv* layer, adding interdependencies information before the final classification. This experiment confronts the outputs from these two sources instead of using the *no neighbors* version implemented in Section 6.4, due to the possibility of comparing classifications produced with the same model, resulting in a more accurate interpretation. The outputs for the auxiliary and final classifiers are represented respectively by gray and black boxes in Figure 6.9.



(a) Friends **(b)** Family

Figure 6.9. Example showing the effects of considering relationships interdependencies, suggesting the model was able to learn couples and families relations structures. (a) The *friends* class is assigned to the correct relationship by the final classifier. (b) The addition of relationships interdependencies allowed the model to adequately identify the *family* relation between the young boy and his father.

For the first example (Figure 6.9a), only the relationships between the woman and the other two individuals are considered. In this case, the auxiliary classifier outputs the same *lovers/spouses* class for both, which is a reasonable classification considering age, gender, proximity, emotion, and other traits taken into account pairwise. However, after considering the information from both relations together, the final classifier changed the output for one of them, suggesting the model was capable of learning that the same person can be involved in only one *lovers/spouses* relationship. Additionally, it selected the correct pair for each class, probably because of their relative distance.

The second sample depicted in Figure 6.9b shows another scenario where the final classifier corrects a previously miss-identified relation. In this case, a *family* class was appropriately assigned to the relationships between the young girl and the other two persons, but the final relation was mistaken by a *commercial* interaction. This probably happened due to the combination of clothing and activity traits from the father, which is a setup similar to the ones seen in *commercial* relations between client and waiter in a restaurant. However, this error was corrected with the addition of information from the sister, allowing the inference of the appropriated classification, as mentioned in Section 2.2.

Finally, Figure 6.10 presents a sample where two relations between separated pairs of persons are considered in the same image. In this situation, the correct *father-child* output is obtained by the auxiliary classifier for both relationships. However, after the addition of interdependencies information by applying the *full neighbors* method, the classification for one of the relations is incorrectly changed to *mother-child*. Similar to the example from Figure 6.9a, this suggests that the model was capable of learning family structures, inferring that if one of the adults is the father, the other must be the mother, which is also a typical setup in various images from both datasets.

This sample is from the PIPA-relation dataset, which does not have a *no relation* label. Additionally, the image is not fully annotated, only providing the ground-truths for two relationships, like many other images from both datasets. In this situation, inserting connections between these pairs has only introduced noise to the decision-making process, as described in Section 5.3. However, the proposed method implementing *social neighbor* connections was able to prevent this issue, producing the correct output for both relationships.

Finally, for some other samples investigated during this experiment, the addition of interdependencies information tended to modify correct outputs for underrepresented classes when they appear simultaneously with multiple other relations sharing a label. For example, if a *lovers/spouses* relationship appears with many other *friends* relations in the same image, the classification can sometimes be changed to *friends* after the addition of interdependencies information. However, this problem is also mitigated by *social neighbor* connections.

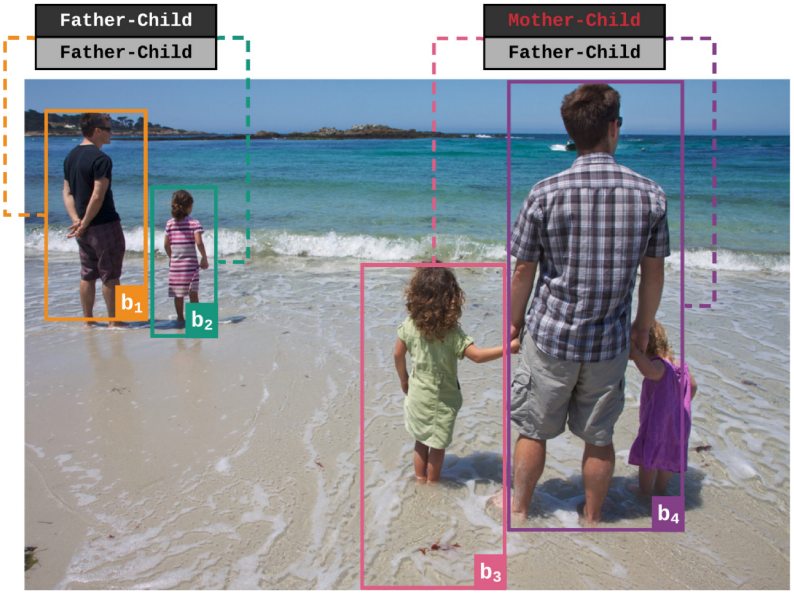


Figure 6.10. Example showing an instance where *full neighbors* connections are applied, adding noise to the model and resulting in a wrong *mother-child* classification, denoted in red. This problem is solved by the proposed *social neighbors* method, preserving the correct *father-child* output from the auxiliary classifier.

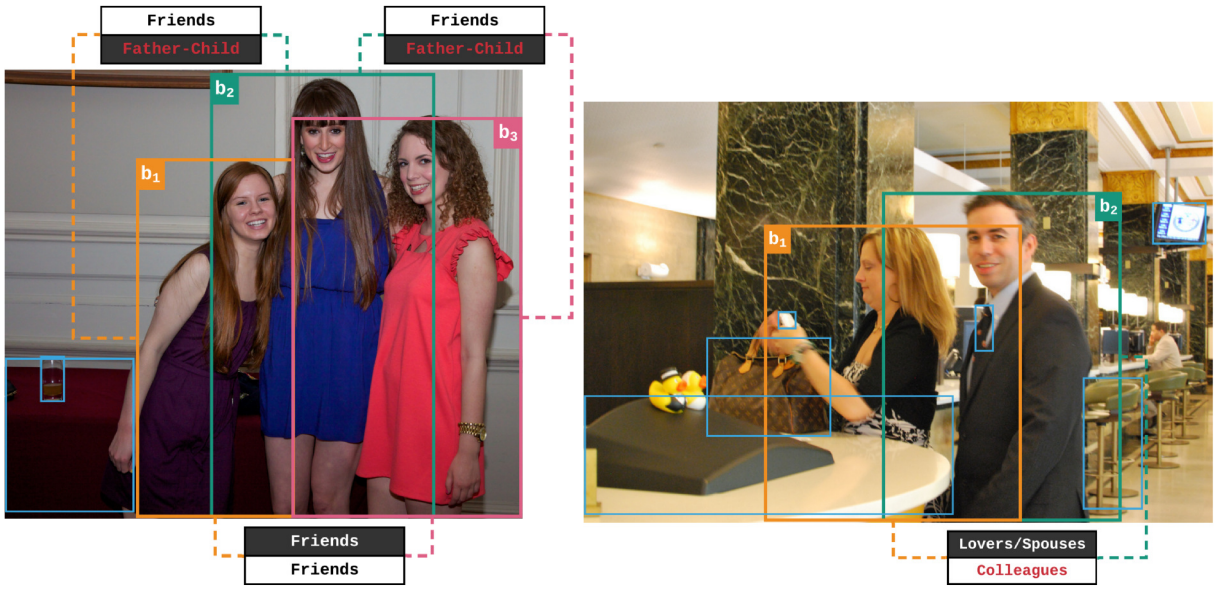
6.5.3 Attributes Features

In this final section, samples classified using the SKG+ version are confronted against the results obtained by the default SKG, providing insight into the influence of attribute features for social relation recognition tasks. The implementation of the extended version of the graph from Section 6.4 was employed for this analysis, including six distinct types of attributes distributed over the three social scales.

Two examples of how the addition of these attribute features can influence the classifications are shown in Figure 6.11, representing the results for the default and extended versions of the graph in black and white, respectively. The bounding boxes of distinct objects detected are also denoted in blue, providing additional information of which object classes are being considered for the image.

Starting with Figure 6.11a, which considers three social relationships, the model using the default graph version incorrectly classified the relationships involving the girl in the middle as *father-child*. This likely happened due to the difference in height with the other girls, which forms a common setup in *parent-child* relationships, as seen from other previous examples. Additionally, the result also suggests that the features learned by the model were unable to adequately abstract gender and age traits for this person, since both are unappropriated to the predicted class.

In this scenario, the addition of attribute features was enough to get adequate results, probably due to personal-scale traits, which included age, gender, and clothing.



(a) Friends **(b)** Colleagues

Figure 6.11. Example of how attribute features affect relationship classifications. (a) Positive sample where the wrong *father-child* output is appropriately changed to *friends*, probably due to age and gender attributes. (b) Negative sample where the correct *lovers/spouses* classification is modified to *colleagues* after the addition of attribute information. The change may be related to formal clothing and the detected objects denoted in blue, which are usually associated with working environments.

The result suggests that this extra information supported the learning process into associating attributes with relationship classes, as described in Section 2.3. More specifically, the *father-child* relationship is discarded after considering gender and age information, and additionally, these cues may have also contributed by increasing the scores for the correct output.

For the second sample, the addition of attribute features was detrimental, causing the appropriated *lovers/spouses* classification to be modified to *colleagues*. In this case, the addition of the clothing attribute may have been the reason, since there is a strong correlation between formal clothing and non-intimate relationships. Additionally, global attributes may also play a role in this result, since the detected object classes included *handbag*, *tie*, and *chair*, which are associated with working environments.

This analysis led to the conclusion that attributes can work in the intended way, constraining the model to learn class-specific dependencies, especially when the dataset has attribute-related labels such as *father-child* and *mother-child*. In these cases, some inappropriate relationship classes can be directly discarded, but these outcomes are strongly dependent on the quality of the attribute features, as mentioned in Section 6.4. Additionally, the models used to obtain this information are required to have a strong generalization capability, allowing the transfer of this knowledge to other datasets. If this is not the case, these attributes may only add noise to the model, resulting in performance loss.

Chapter 7

Conclusion

Current social relation recognition methods are focused on specific aspects of the problem, which are treated separately. However, the dependencies between information from multiple sources and different scales form an structure, which is fundamental to classify social relationships adequately. These approaches are based on inaccurate interpretations of relationships, resulting in models that are unable to capture the full context necessary to identify social relations.

In this context, a new approach to interpreting image-based social relation recognition methods was introduced in this work, examining three central aspects: the scope of the employed information, the capacity to apply prior knowledge on each of these sources, and the consideration of their interdependencies within an image. These criteria were used to evaluate previous works, identifying their strengths and shortcomings, and this knowledge was applied to develop a new methodology.

The proposed approach introduced the Social Knowledge Graph (SKG), a representation for social relationships capable of achieving information completeness by preserving the structure of the relationships from the input image. This means that it can represent learned features, pre-trained attributes, along with other types of prior knowledge and constraints associated with three distinct information scopes, which are denominated in this work as social scales. More precisely, the graph carries individual traits from the personal scale, pairwise relative information from the local scale, and general context features from the global scale. Each of these scopes offers a unique point of view over social interactions, contributing with meaningful complementary data.

All this information is gathered through the proposed Social Scales Network (SSN), which assigns a convolutional backbone to learn features from individual scales separately. This approach allows each model to specialize in aspects that are relevant to their specific scopes, capturing data with different granularities. No previous work was capable of extracting information from all these sources simultaneously, producing incomplete social relationship models. Another major benefit of this approach is retaining the relationship structure within an image, which allows the consideration of multi-scale dependencies between learned features, pre-trained attributes, and different types of prior knowledge from other relationships within an image.

A deep graph model was proposed to exploit the rich and intuitive SKG representation, employing a message passing mechanism explicitly designed to take advantage of the intuitive structure. The Social Graph Network (SGN) is composed of three layers, where each one is responsible for leveraging information embedded in specific regions of the graph. The first layer aggregates multi-attribute information for all scales, which are combined with features learned directly from the input image using their respective scopes. The second one generates multi-scale representations for the relationships in the image by combing the features from all social scopes, and the last layer is in charge of capturing the interdependencies between all the information extracted by other relationships. A specific spatial convolution operation was proposed for each of these layers, based on the region of the graph they are applied, optimizing the feature aggregation process according to the information type. Finally, the noise carried by all this information is reduced with the help of graph-level structures and a reasoning-level attention mechanism.

Multiple experiments were conducted to evaluate the proposed framework, generating results that surpassed all previous methods, achieving a new state-of-the-art. A rigorous quantitative analysis was conducted by evaluating the effects of model variations and estimating the significance of individual modules through an ablation study, highlighting the significance of each part. The results indicate the validity of the proposed interpretation of social relationships, especially when considering the role of multi-scale features, prior knowledge, and relationships interdependencies to the final model performance.

Lastly, a qualitative analysis was presented and discussed, showing examples of the interactions between information from different scales, the role of pre-defined attributes, and the effects of relationship interdependencies on the model's decision-making process. The experiments indicated that the proposed concepts worked as intended, providing reasonable outcomes and generating predictions in line with social theories, bringing automated relation recognition closer to human perception of social relationships.

Chapter 8

Future Work

Future work can investigate the effects of adding new information to the SKG, such as different attributes and additional prior knowledge to each scale. Another possibility is learning these features jointly with the social relation recognition task, which can help with the generalization capabilities of the model. The main challenge of this approach would be the lack of databases that simultaneously offer these types of information. However, this problem could be circumvented with the use of oracle networks, or maybe with the construction of such a dataset, which would be a significant contribution.

The proposed Social Knowledge Graph does not carry any information in its edges, which serves only as a differentiation between data types. In this sense, extra features could be added to the edges, representing relative information regarding the nodes they connect. For example, relation edges could carry the distance between their corresponding pairs, also establishing relative information between relationships, and not only individuals, as is currently done. These adaptations have to be reflected in the proposed aggregation methods since they also do not consider edge features.

New convolution operations, message passing mechanisms, and other technical improvements to the SGN are also relevant research directions. These kinds of contributions have the advantage of being extensible to multiple other tasks. However, they require a more profound knowledge of graph neural networks, which can be highly time-consuming and can also be considered as out of the scope of the social relation recognition problem.

A final potential research direction would be to extend the SKG and the SGN to the temporal domain, allowing the method to be employed in social relation recognition from video. This could be done by generating graph representations for a sequence of frames connected with temporal edges. The convolution operations also have to be adapted to consider this new structure.

Bibliography

Aimar, E. S., Radeva, P., and Dimiccoli, M. (2019). Social relation recognition in egocentric photostreams. *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3227–3231.

Alecrim, E. (2015). Pepper, o robô que "lê" emoções e teve mil unidades vendidas em um minuto no Japão. `https://tecnoblog.net/180170/robo-pepper-japao/`. [Online; accessed 03-November-2021].

August, K. J. and Rook, K. S. (2013). *Social Relationships*. Springer, New York, NY. ISBN 978-1-4419-1005-9.

Barkan, S. E. (2011). *Sociology: Understanding and Changing the Social World*. Open Textbooks, Minneapolis, MN.

Bartlett, M., Edmunds, C., Belpaeme, T., Thill, S., and Lemaignan, S. (2019). What can you see? identifying cues on internal states from the movements of natural social interactions. *Front. Robotics and AI*, 2019.

Brooks, R. (2017). The Big Problem With Self-Driving Cars Is People. `https://spectrum.ieee.org/transportation/self-driving/the-big-problem-with-selfdriving-cars-is-people`. [Online; accessed 10-June-2019].

Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. (2014). Spectral networks and locally connected networks on graphs. *Computing Research Repository (CoRR)*, abs/1312.6203.

Buades, A., Coll, B., and Morel, J. (2005). A non-local algorithm for image denoising. *2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2:60–65 vol. 2.

Bugental, D. (2000). Acquisition of the algorithms of social life: A domain-based approach. *Psychological bulletin*, 126:187–219.

Cao, S., Lu, W., and Xu, Q. (2016). Deep neural networks for learning graph representations. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 1145–1152. AAAI Press.

Casagrande, V. (2019). Gol lança robô para atendimento ao passageiro no aeroporto de Guarulhos. https://economia.uol.com.br/todos-a-bordo/2019/10/22/gol-lanca-robo-para-atendimento-ao-passageiro-no-aeroporto-de-guarulhos.htm. [Online; accessed 11-October-2019].

Chen, C., Zhang, R., Koh, E., Kim, S., Cohen, S., and Rossi, R. (2020). Figure captioning with relation maps for reasoning. In *IEEE Winter Conference on Applications of Computer Vision*.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724--1734, Doha, Qatar. Association for Computational Linguistics.

Dai, P., Lv, J., and Wu, B. (2019). Two-stage model for social relationship understanding from videos. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1132–1137.

Dai, Q., Carr, P., Sigal, L., and Hoiem, D. (2015). Family member identification from photo collections. *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 982–989.

Daraya, V. (2013). Robô Link237 recepciona clientes do Bradesco em SP. https://exame.abril.com.br/tecnologia/robo-link237-recepciona-clientes-do-bradesco-em-sp/2/. [Online; accessed 22-November-2019].

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3844–3852, Red Hook, NY, USA. Curran Associates Inc.

Dhall, A., Joshi, J., Sikka, K., Goecke, R., and Sebe, N. (2015). The more the merrier: Analysing the affect of a group of people in images. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 1:1–8.

Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2224–2232, Cambridge, MA, USA. MIT Press.

Fairclough, N., Corporation, E., Routledge, and dawsonera (2003). *Analysing Discourse: Textual Analysis for Social Research*. Analysing Discourse: Textual Analysis for Social Research. Routledge. ISBN 9780415258920.

Fiske, A. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological review*, 99:689–723.

Gallicchio, C. and Micheli, A. (2010). Graph echo state networks. *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Gemeren, C. V., Poppe, R., and Veltkamp, R. C. (2018). Hands-on: deformable pose and motion models for spatiotemporal localization of fine-grained dyadic interactions. *EURASIP Journal on Image and Video Processing*, 2018:1–16.

Ghosh, S., Dhall, A., and Sebe, N. (2018). Predicting group cohesiveness in images. *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1263–1272. Journal of Machine Learning Research (JMLR).

Girshick, R. B. (2015). Fast r-cnn. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249--256, Chia Laguna Resort, Sardinia, Italy. PMLR.

Goel, A., Ma, K. T., and Tan, C. (2019). An end-to-end network for generating social relationship graphs. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11178–11187.

Gori, M., Monfardini, G., and Scarselli, F. (2005). A new model for learning in graph domains. In *IEEE International Joint Conference on Neural Networks.*, volume 2, pages 729–734 vol. 2.

Guo, X., Polania, L., Zhu, B., Boncelet, C., and Barner, K. (2020). Graph neural networks for image understanding based on multiple cues: Group emotion recognition and event recognition as use cases. In *IEEE Winter Conference on Applications of Computer Vision*.

Guo, X., Polanía, L. F., Garcia-Frias, J., and Barner, K. E. (2019). Social relationship recognition based on a hybrid deep neural network. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, pages 1–5.

Guo, X., Zhu, B., Polanía, L. F., Boncelet, C., and Barner, K. E. (2018). Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*.

Hassaballah, M., Abdelmgeid, A. A., and Alshazly, H. A. (2016). *Image Features Detection, Description and Matching*. Springer International Publishing, Cham. ISBN 978-3-319-28854-3.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. B. (2020). Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:386–397.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Henaff, M., Bruna, J., and LeCun, Y. (2015). Deep convolutional networks on graph-structured data. *ArXiv*, abs/1506.05163. [Online; accessed 14-July-2021].

Herath, S., Harandi, M., and Porikli, F. (2017). Going deeper into action recognition: A survey. *Image and Vision Computing*, 60:4–21. ISSN 0262-8856. Regularization Techniques for High-Dimensional Data Analysis.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9:1735–1780.

Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E. (2020). Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:2011–2023.

Ibrahim, M. S., Muralidharan, S., Deng, Z., Vahdat, A., and Mori, G. (2015). A hierarchical deep temporal model for group activity recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1971–1980.

Ilemo, S. N., Barth, D., David, O., Quessette, F., Weisser, M.-A., and Watel, D. (2019). Improving graphs of cycles approach to structural similarity of molecules. *PLoS ONE*, 14.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Bach, F. R. and Blei, D. M., editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France,*

6-11 July 2015, volume 37 of *Journal of Machine Learning Research (JMLR) Workshop and Conference Proceedings*, pages 448--456. JMLR.org.

Jain, A., Zamir, A., Savarese, S., and Saxena, A. (2016). Structural-rnn: Deep learning on spatio-temporal graphs. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5308–5317.

Johnson, J., Krishna, R., Stark, M. A., Li, L., Shamma, D., Bernstein, M. S., and Fei-Fei, L. (2015). Image retrieval using scene graphs. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678.

Kiesler, D. (1983). The 1982 interpersonal circle: A taxonomy for complementarity in human transactions. *Psychological Review*, 90:185–214.

Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

Ko, B. (2018). A brief review of facial emotion recognition based on visual information. In *Sensors*.

Kong, Y. and Fu, Y. (2018). Human action recognition and prediction: A survey. *ArXiv*, abs/1806.11230. [Online; accessed 14-July-2021].

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., and Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *International Journal of Computer Vision*, page 32–73.

Lan, T., Wang, Y., Yang, W., Robinovitch, S. N., and Mori, G. (2012). Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:1549–1562.

Lei Ba, J., Kiros, J. R., and Hinton, G. E. (2016). Layer Normalization. *arXiv e-prints*, page arXiv:1607.06450. [Online; accessed 14-July-2021].

Levie, R., Monti, F., Bresson, X., and Bronstein, M. (2019). Cayleynets: Graph convolutional neural networks with complex rational spectral filters. *IEEE Transactions on Signal Processing*, 67:97–109.

Li, J., Wong, Y., Zhao, Q., and Kankanhalli, M. S. (2017). Dual-Glance Model for Deciphering Social Relationships. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2669–2678.

Li, T., Chang, H., Wang, M., Ni, B., Hong, R., and Yan, S. (2015). Crowded scene analysis: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(3):367–386. ISSN .

Li, W., Duan, Y., Lu, J., Feng, J., and Zhou, J. (2020). Graph-based social relation reasoning. In *The European Conference on Computer Vision (ECCV)*.

Li, Y., Tarlow, D., Brockschmidt, M., and Zemel, R. S. (2016). Gated graph sequence neural networks. In Bengio, Y. and LeCun, Y., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Li, Y., Yu, R., Shahabi, C., and Liu, Y. (2018). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv: Learning*. [Online; accessed 14-July-2021].

Lindeberg, T. (1994). Scale-space theory : A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, 21:225–270.

Liu, X., Liu, W., Zhang, M., Chen, J., Gao, L., Yan, C., and Mei, T. (2019). Social relation recognition from videos via multi-scale spatial-temporal reasoning. In *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Loshchilov, I. and Hutter, F. (2019). Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Lu, C., Krishna, R., Bernstein, M., and Fei-Fei, L. (2016). Visual Relationship Detection with Language Priors. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 852--869, Cham. Springer International Publishing.

Lv, J., Liu, W., Zhou, L., Wu, B., and Ma, H. (2018). Multi-stream fusion model for social relation recognition from videos. In *MultiMedia Modeling (MMM)*.

LV, J., WU, B., ZHANG, Y., and XIAO, Y. (2019). Attentive sequences recurrent network for social relation recognition from video. *IEICE Transactions on Information and Systems*, E102.D(12):2568–2576.

Meng, Z., Adluru, N., Kim, H. J., Fung, G., and Singh, V. (2018). Efficient relative attribute learning using graph neural networks. In *The European Conference on Computer Vision (ECCV)*.

Micheli, A. (2009). Neural network for graphs: A contextual constructive approach. *IEEE Transactions on Neural Networks*, 20:498–511.

Niepert, M., Ahmed, M., and Kutzkov, K. (2016). Learning convolutional neural networks for graphs. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 2014–2023. Journal of Machine Learning Research (JMLR).

Noceti, N. and Odone, F. (2014). Humans in groups: The importance of contextual information for understanding collective activities. *Pattern Recognit.*, 47:3535–3551.

Oh, S. J., Benenson, R., Fritz, M., and Schiele, B. (2020). Person recognition in personal photo collections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:203–220.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In *The British Machine Vision Conference (BMVC)*.

Raboh, M., Herzig, R., Berant, J., Chechik, G., and Globerson, A. (2020). Differentiable scene graphs. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Ramanathan, V., Yao, B., and Fei-Fei, L. (2013). Social role discovery in human events. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2475–2482.

Ren, S., He, K., Girshick, R. B., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*, 39:1137–1149.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *ArXiv*, abs/1706.05098. [Online; accessed 14-July-2021].

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2009). The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

Seo, Y., Defferrard, M., Vandergheynst, P., and Bresson, X. (2018). Structured sequence modeling with graph convolutional recurrent networks. *International Conference on Neural Information Processing*, pages 362–373.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *Computing Research Repository (CoRR)*, abs/1409.1556.

Singla, P., Kautz, H. A., Luo, J., and Gallagher, A. (2008). Discovery of social relationships in consumer photo collections using markov logic. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–7.

Smith, E. and Zárate, M. (1990). Exemplar and prototype use in social categorization. *Social Cognition*, 8:243–262.

Sperduti, A. and Starita, A. (1997). Supervised neural networks for the classification of structures. *IEEE transactions on neural networks*, 8 3:714–35.

Sun, Q., Schiele, B., and Fritz, M. (2017). A Domain Based Approach to Social Relation Recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 435–444.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

Teney, D., Liu, L., and van den Hengel, A. (2017). Graph-structured representations for visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Thomee, B., Shamma, D., Friedland, G., Elizalde, B., Ni, K. S., Poland, D. N., Borth, D., and Li, L. (2016). Yfcc100m: the new data in multimedia research. *Commun. ACM*, 59:64–73.

Varadarajan, J., Subramanian, R., Bulò, S. R., Ahuja, N., Lanz, O., and Ricci, E. (2017). Joint estimation of human pose and conversational groups from social scenes. *International Journal of Computer Vision*, pages 1--20.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph Attention Networks. *International Conference on Learning Representations*.

Wang, D., Cui, P., and Zhu, W. (2016a). Structural deep network embedding. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Gool, L. V. (2016b). Temporal segment networks: Towards good practices for deep action recognition. *Computing Research Repository (CoRR)*, abs/1608.00859. [Online; accessed 14-July-2021].

Wang, M., Du, X., Shu, X., Wang, X., and Tang, J. (2020). Deep supervised feature selection for social relationship recognition. *Pattern Recognit. Lett.*, 138:410–416.

Wang, X., Girshick, R. B., Gupta, A., and He, K. (2018a). Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803.

Wang, Z., Chen, T., Ren, J., Yu, W., Cheng, H., and Lin, L. (2018b). Deep Reasoning with Knowledge Graph for Social Relationship Understanding. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1021--1028. International Joint Conferences on Artificial Intelligence Organization.

Wu, J., Wang, L., Wang, L., Guo, J., and Wu, G. (2019). Learning actor relation graphs for group activity recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32:4–24.

Xiao, B., Wu, H., and Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In *The European Conference on Computer Vision (ECCV)*.

Yan, H. and Song, C. (2019). Semantic three-stream network for social relation recognition. *Pattern Recognition Letters*, 128:78 – 84. ISSN 0167-8655.

Yan, X., Kakadiaris, I., and Shah, S. (2014). Modeling local behavior for predicting social interactions towards human tracking. *Pattern Recognit.*, 47:1626–1641.

Yatskar, M., Zettlemoyer, L., and Farhadi, A. (2016). Situation recognition: Visual semantic role labeling for image understanding. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5534–5542.

Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., and Leskovec, J. (2018). Graph convolutional neural networks for web-scale recommender systems. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Zhang, M., Liu, X., Liu, W., Zhou, A., Ma, H., and Mei, T. (2019). Multi-Granularity Reasoning for Social Relation Recognition From Images. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1618–1623.

Zhang, N., Paluri, M., Taigman, Y., Fergus, R., and Bourdev, L. (2015). Beyond frontal faces: Improving Person Recognition using multiple cues. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4804–4813.

Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2015). Learning social relation traits from face images. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3631–3639.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81. ISSN 2666-6510.