AMERICAN UNIVERSITY OF BEIRUT

HUMAN OBJECT INTERACTION DETECTION IN PAINTINGS USING MULTI-TASK LEARNING

by
MAYA JACK ANTOUN

A dissertation
submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
to the Department of Mechanical Engineering
of Maroun Semaan Faculty of Engineering and Architecture
at the American University of Beirut

Beirut, Lebanon
June 2023

AMERICAN UNIVERSITY OF BEIRUT

HUMAN OBJECT INTERACTION DETECTION IN PAINTINGS USING MULTI-TASK LEARNING

by

MAYA JACK ANTOUN

Approved by:

Dr. Daniel Asmar, Professor Mechanical Engineering	Advisor
Dr. Kamel Abu Ghali, Professor Mechanical Engineering	Chairperson of Committee
Dr. Imad H. ElHajj, Professor Electrical and Computer Engineering	Member of Committee
Dr. Joe Tekli, Associate Professor Electrical and Computer Engineering, Lebanese American University	Member of Committee
Dr. Najib Metni, Associate Professor Mechanical Engineering, Notre-Dame University	Member of Committee

Date of dissertation defense: June 20, 2023

AMERICAN UNIVERSITY OF BEIRUT

DISSERTATION RELEASE FORM

Student Name:	Antoun	Maya	Jack
	Last	First	Middle

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of my dissertation; (b) include such copies in the archives and digital repositories of the University; and (c)make freely available such copies to third parties for research or educational purposes

_**X**_ **As of the date of submission of my dissertation.**

___ **After 1 year from the date of submission of my dissertation.**

--- **After 2 years from the date of submission of my dissertation.**

--- **After 3 years from the date of submission of my dissertation.**

06 July 2023

Signature Date

Acknowledgements

I would like to take a moment to extend my heartfelt gratitude to my advisor, Prof. Daniel Asmar, for his guidance throughout my thesis journey. Prof. Asmar's invaluable expertise and insights have been instrumental in shaping the direction of my research and in pushing me to achieve my academic goals. I have known Dr. Asmar for more than a decade, and throughout this time, he has always been a source of inspiration and motivation for me. Even during my lowest points, you were there to offer words of encouragement, reminding me of the importance of perseverance and dedication in achieving success. I would also like to thank all the committee members, Prof. Abu Ghali,, Prof. ElHajj,, Prof. Tekli, and Prof. Metni for their valuable time, feedback, and advice. Your comments and suggestions had a significant contribution in completing my thesis objectives. I would also like to extend my sincere thanks to the AUB University Research Board (URB) and the National Council for Scientific Research of Lebanon (CNRS-L) for their generous funding and support of my research.

I would like to also thank to my parents and family for their unwavering love, motivation, and support throughout my thesis journey. Their belief in me has been my greatest strength. My parents have been my pillars of strength, both in my personal and professional life. They played a crucial role in enabling me to balance my academic and personal responsibilities. You always took care of me and my daughter when I was busy with work and meetings, allowing me to focus on my research without worrying about other responsibilities. Your support and understanding have been invaluable, and I am deeply grateful for all that you have done for me.

To my husband Oliver and my precious daughter Jude, I owe it all to you. Oliver, you have been my constant support, providing me with a listening ear and words of encouragement whenever I needed them. You are my rock, always pushing me to keep going even when I felt like giving up. I cannot thank you enough for staying up late with me, rehearsing my presentations and being there for me every step of the way. Your support and love have been the driving force behind my success, and I am more than grateful to have you in my life. Jude, you are the reason I did not give up when I wanted to. Being a mother and a role model to you has been one of the most fulfilling experiences of my life, and I hope that this achievement will inspire you to pursue your own dreams and aspirations in life. I am so proud of you, my dear daughter, and I hope that you will always be proud of me too.

Abstract
of the Dissertation of

Maya Jack Antoun for Doctor of Philosophy
Major: Mechanical Engineering

Title: Human Object Interaction Detection in Paintings using Multi-Task Learning

Human Object Interaction (HOI) detection provides valuable insights into the meaning and interpretation of a painting, as the interactions between humans and object reveal information about the scene, characters, and story depicted in the artwork. Automatically detecting HOI in paintings is a challenging task, as the paintings often contain complex scenes with intricate details and variations in artistic style. Additionally, unlike in real-world images, the context and physics of the painting may not follow physical rules, which can further complicate the detection process.

The proposed system addresses the complexities of this task, considering the intricate details and variations in artistic style found in paintings. It incorporates a model that captures discriminative information by extracting visual features from detected humans, objects, and the Region of Interest. The model analyzes spatial arrangements to understand the relationships and interactions between elements. Moreover, the model integrates contextual knowledge and semantic relationships using a knowledge graph based on Graph Convolution Network to capture the underlying meaning and story depicted in artwork.

However, relying solely on appearance and context may not be enough to accurately infer HOIs in paintings. To overcome this challenge, multitask learning is employed by introducing four supplementary classification tasks. These tasks provide complementary information that enhances the HOI detection process, leveraging shared representations across multiple tasks. The proposed system introduces the SemArt-HOI benchmark dataset, augmenting the SemArt dataset with instance detection annotations and interaction classes. Experimental results demonstrate that the proposed model outperforms the state-of-the-art one-stage transformer-based HOI detection model in both single-task and multi-task settings by 1.19% and 1.51% respectively. Furthermore, the system exhibits superior efficiency, training four times faster and requiring fewer resources. This makes it suitable for practical and large-scale HOI detection in paintings.

Table of Contents

Illustrations

Tables

Abbreviations

AP	Average Precision
BERT	Bidirectional Encoder Representations from Transformers
biLM	bidirectional language model
CAGrad	Conflict-Averse Gradient Descent
CNN	Convolutional Neural Network
CPN	Cascaded Pyramid Network
DETR	DEtection TRansformer
DIRV	Dense Interaction Region Voting
DITCH	DIgiTal Cultural Heritage center
DT	Default
DWA	Dynamic Weight Average
ELMo	Embeddings from Language Models)
EMA	Exponential Moving Average
FP	False Positive
FCN	Fully Connect Network
FFN	Feed Forward Networks
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GIST	Generalized Search Trees
GloVe	Global Vectors for Word Representation
GradNorm	Gradient Normalization
GNN	Graph Neural Network
HAG	Hierarchical Activity Graph
HICO	Humans Interacting with Common Objects
HOA	Higher Order Approximation
HOI	Human Object Interaction
IOU	Intersection Over Union
KO	Known-Object
LBP	Local Binary Patterns
mAP	mean Average Precision
MET	METropolitan Museum of New York
MLP	Multi-Layer Perceptron
MTL	Multi-Task Learning
NLP	Natural Language Processing
PaSta	Human Body Part States

ResNet	Residual Neural Network
ROI	Region Of Interest
RLW	Random Loss Weight
RMPE	Regional Multi-person Pose Estimation
SGD	Stochastic Gradient Descent
SemArt	Semantic Art
STL	Single-Task Learning
STN	Spatial Transformer Networks
TP	True Positive
UW	Uncertainty Weights
V-COCO	Verbs in COCO
ViT	Vision Transformer
2TL	2-Task Learning
3TL	3-Task Learning
4TL	4-Task Learning
5TL	5-Task Learning

Chapter 1

Introduction

Scene understanding in paintings is a challenging task that aims to analyze the various features and elements within a painting to retrieve its multiple attributes. One of the most important features is identifying the style and genre of the painting. Different art movements have their own distinctive features, and understanding these can help to identify the artist and the time period during which the painting was created. Recognizing the artist behind the painting is also a critical task in scene understanding, as it can provide important insights into their influences, techniques, and overall artistic style.

1.1 Background and Rationale of the Thesis

An important aspect of scene understanding in paintings is the analysis of the composition of the painting, particularly the presence of humans and objects, and the relationships between them. This involves identifying and understanding the roles of different characters within the painting, the emotions and expressions on their faces, and their interactions with other objects and humans in the scene. For example, in a painting depicting a battlefield, it is important to identify the different soldiers, commanders, and other figures in the painting, and their relationships to one another. This could include analyzing the expressions on their faces to understand their emotions, as well as the placement of objects within the scene to understand the overall composition and story being told. By analyzing the presence of humans and objects in paintings and their relationships to one another, we can gain a deeper understanding of the painting's meaning and the artist's intention behind it. This type of analysis can be particularly useful in art history and cultural studies, as well as in fields such as computer vision and machine learning where scene understanding is a key challenge.

Compared to scene understanding in natural images, the task of understanding an artistic representation in paintings is undoubtedly more complex. In natural images, object detection and recognition rely on the visual appearance and context in the image. However, in paintings, identifying the elements present requires an understanding of the symbolism, metaphors, and artistic conventions used by the artist. For example, recognizing the significance of a particular color or object within

a painting may require knowledge of the cultural or historical context in which the artwork was created. Understanding the roles of different characters in a painting, the emotions and expressions on their faces, and their interactions with other objects and humans in the scene are crucial for scene understanding in paintings. It requires an understanding of the social, cultural, and historical context in which the painting was created. Overall, scene understanding in paintings involves multiple complex processes that require a multidisciplinary approach combining visual perception, art history, and cultural studies.

Psychologically, the challenge of detecting human-object interactions in paintings can be attributed to the concept of 'cognitive schema' - mental frameworks that help individuals organize and interpret information about the world around them. Cognitive schema [1], [2] are developed through experience and learning, and they allow individuals to make predictions and inferences based on their past experiences. However, when presented with information that does not fit into their existing schema, individuals may struggle to interpret and make sense of the new information. In the context of human object interaction detection in paintings, the violation of physical rules and the intentional manipulation of context by artists can create situations where traditional cognitive schema may not be applicable, making it more challenging to detect and interpret human-object interactions.

Despite these challenges, my thesis addresses the task of human-object interaction detection in paintings, aiming to develop methodologies and techniques that overcome the limitations imposed by cognitive schema and the artistic context. The following are the main contributions of my research:

- I have developed the first human object interaction detection system specifically designed for paintings. This system effectively detects and localizes both humans and objects present in paintings, along with capturing their interactions. The system operates in two stages. In the first stage, I fine-tune a CNN-based object detector to accurately identify all potential instances of human-object interaction within the artwork. This initial stage serves as a crucial foundation for subsequent analysis. In the second stage, I extract visual, spatial, and semantic features from the detected instances. These features are then integrated and fused together to predict the specific human-object interactions in the painting. By combining both visual and spatial information, the system achieves a comprehensive understanding of the interactions between humans and objects in paintings, and by using semantic features the model gains a deeper understanding of the meaning of the paintings thereby facilitating deeper insights into the artistic representation and narrative. No pose features were included in the proposed system because in paintings, the human pose can be misleading and may not accurately represent the intended interaction between the human and object.

- My system undergoes improvement through the utilization of a multi-task learning approach, which involves incorporating tasks related to type, school, timeframe, and author information. This approach enhances the second stage of interaction prediction by introducing four additional output prediction tasks.

Consequently, the model is trained to classify a total of five tasks simultaneously. The incorporation of these extra tasks enables the model to leverage the shared representation of features, resulting in enhanced performance in predicting interactions. This approach effectively harnesses the benefits of multi-task learning, allowing the model to gain a broader understanding of the painting data and improve its ability to predict human-object interactions.

- I introduce the SemArt-HOI dataset, the first dataset specifically curated for human object interaction detection in paintings. This dataset builds upon the existing SemArt dataset, which includes classification labels for various attributes such as painting type, school, timeframe, and author. SemArt-HOI expands upon this foundation by incorporating object detections and interaction labels, thus enriching the dataset with valuable information for HOI analysis. By combining the SemArt classification labels with object detections and interaction annotations, SemArt-HOI becomes a comprehensive resource for training and evaluating HOI detection models specifically tailored for paintings.

- In the conducted experiments using the SemArt-HOI dataset, I thoroughly evaluated the performance of my proposed approach for human object interaction (HOI) detection in paintings. The evaluation involved comparing the performance of my approach with the existing state-of-the-art systems in this field. The results of the evaluation demonstrated that my approach outperformed the current state-of-the-art system significantly. These results highlight the effectiveness and superiority of my proposed approach in the context of HOI detection in paintings. By surpassing the performance of the state-of-the-art system on the SemArt-HOI dataset, my approach opens up new possibilities for advancing the understanding and analysis of human-object interactions in artistic representations.

- In addition to its superior performance, my system demonstrates faster training compared to the state-of-the-art system on the SemArt-HOI dataset. Specifically, my CNN-based model exhibits faster training speed compared to the transformer-based approach, and it also requires less GPU usage, thereby enhancing its efficiency. The accelerated training time holds significant advantages, particularly in practical applications that demand real-time or near real-time performance. The reduced training time not only increases the efficiency of the system but also enhances its usability in scenarios where quick results are essential. By significantly decreasing the time required for training, my system enables faster iteration and experimentation, facilitating the development of more accurate and robust models for human object interaction detection in paintings.

Chapter 2

Scene Understanding in Paintings

Scene understanding involves analyzing objects in their context, taking into account the scene's structure, layout, and the spatial, functional, and semantic relationships between objects. Even with a quick glance, one can make a general classification of a scene, such as identifying it as an outdoor park or an indoor theater. Additionally, certain categories of objects, people, and animals can be recognized even from brief exposures. By examining the features of a scene, it is possible to infer properties of these objects, such as a person's gender and potentially their emotions. Actions can also be characterized by their distinct features. Therefore, beyond simply detecting the presence of people in a scene, one can also extract basic information about them, including their gender, emotions, and the actions they are performing. This comprehensive understanding of a scene goes beyond object detection and allows for a more nuanced interpretation of the elements within it. By analyzing the scene's composition and the characteristics of its objects, we gain valuable insights into the scene's dynamics and the relationships between its components.

2.1 Digitization of Painting

Throughout history, humans have utilized paintings as a means of expressing their emotions and providing insights into their way of life. Paintings offer a visual representation of the past that surpasses written words, granting us a glimpse into different eras and fostering an understanding of how people existed during those times. For example, artists have captured significant battles and historical events on various mediums, be it paper or canvas. Paintings possess the ability to preserve memories and immortalize moments, much like contemporary photographs taken with cameras. By observing paintings, we can delve deeper into our cultural heritage and comprehend the progression of human civilization throughout the ages. This rich tradition traces back to cave drawings, which are regarded as one of the earliest forms of communication among people.

In 2017, archaeologists made a significant discovery in Indonesia, uncovering the oldest known painting in history (Figure 2.1). It was found within the depths of

the Leang Tedongnge cave and showcased a depiction of three wild pigs, painted approximately 45,500 years ago on the Indonesian island of Sulawesi, deepening our understanding of our artistic roots. The painting provides valuable insights into the life and culture of hunter-gatherer societies during that time. It depicts a warty pig, highlighting its significance in the daily lives of these communities. The portrayal of warty pigs in Ice Age rock art not only serves as a representation of the animal itself but also carries deeper symbolic and potentially spiritual meanings within the ancient hunting culture of Sulawesi. These paintings offer clues about the importance and reverence given to warty pigs, shedding light on the beliefs, rituals, and societal dynamics of the people who created these artworks. By studying these historical paintings, researchers and historians can gain a better understanding of the past and the cultural significance of animals in the lives of ancient societies.

Figure 2.1: The painting of a wild pig in the Leang Tedongnge cave on the Indonesian island of Sulawesi. Credit: Maxime Aubert

People began recognizing the importance of preserving and showcasing paintings in museums, leading to the establishment of renowned institutions like the Louvre Museum in Paris, France, which offers visitors a chance to immerse themselves in a world of artistic excellence. With a vast collection of over 38,000 artworks, including an impressive selection of 5,500 captivating paintings, the Louvre is a treasure trove of artistic masterpieces. Within its walls, art enthusiasts can marvel at the timeless creations crafted by legendary artists such as Michelangelo, Raphael, and Leonardo da Vinci. Similarly, the Metropolitan Museum of Art in New York City, USA, stands as another notable institution dedicated to the world of art. This prestigious museum houses a rich and extensive compilation of over 6,000 artistic works. Among its remarkable collection are 2,500 paintings created by prominent figures in art history, including Claude Monet, Vincent Van Gogh, and Pablo Picasso. The Metropolitan Museum of Art serves as a cultural beacon, providing visitors with an opportunity

to appreciate and explore the diverse range of artistic expressions that have shaped our heritage.

Both the Louvre Museum and the Metropolitan Museum of Art play pivotal roles in preserving and showcasing these exceptional works of art. By creating spaces for public access and appreciation, these museums contribute to the collective understanding and appreciation of art's cultural significance. Through their efforts, they continue to inspire and educate visitors, ensuring that these masterpieces endure as a testament to human creativity and ingenuity.

The digitization of paintings has revolutionized the accessibility and exploration of art, allowing art enthusiasts to engage with vast collections from the comfort of their own homes. Online platforms like Artsy and WikiArt have emerged as prominent digital galleries, offering diverse collections and interactive features. Artsy (https://www.artsy.net/), founded in 2009, has become the world's largest online art marketplace. It hosts over 1 million artworks from more than 4,000 galleries and top auction marketplaces. With a wide range of mediums including paintings, sculptures, and films, Artsy provides a digital space for users to discover, appreciate, and even buy or sell artwork. Its collection features works from over 100,000 artists, both established and emerging. WikiArt (http://www.wikiart.org), formerly known as WikiPaintings, is an online encyclopedia of visual art launched in 2010. It offers a platform for users to contribute and edit content, fostering a collaborative environment. Additionally, WikiArt facilitates online shopping of art reproductions. Its collection encompasses approximately 250,000 artworks from over 100 countries, including pieces displayed in various institutions such as museums, universities, and civic buildings.

As the digital collections of artworks grow, the development of multimedia systems for archiving and retrieval becomes essential. These collections often come with metadata in the form of annotations or tags provided by art historians and curators. These annotations contain valuable information about the artist, style, date, genre, school, and other details related to each painting. To facilitate searching for specific artworks, these tags, associated with scene understanding tasks, are utilized as search filters. In Artsy, users can search for artworks using various criteria such as the artist's name, nationality or ethnicity, the material used, the gallery or city of origin, the time period, or even specific colors present in the artwork. This wide range of search options allows users to tailor their exploration based on their preferences and interests. Similarly, in WikiArt, users can search for artworks based on the artist's name, nationality, school, art movement, or the specific centuries during which the painting was created. Furthermore, WikiArt offers additional search filters including 229 different styles, 68 genres, and 244 painting media. These comprehensive search options in WikiArt enable users to delve deeper into specific artistic styles, genres, or mediums, enhancing their exploration and discovery of artworks.

The digitization of paintings and the availability of online platforms like Artsy and WikiArt have significantly expanded the reach and accessibility of art, providing art enthusiasts with unprecedented opportunities to engage with diverse collections and deepen their understanding and appreciation of artistic creations.

2.2 HOI in Paintings

Human Object Interaction (HOI) is an emerging area in image understanding that focuses on detecting and recognizing the relationships between humans and objects in a scene. When a person views an image, they first try to localize and recognize each object in the scene, asking themselves "What is Where?" before trying to understand the relationships between them, asking "What is Happening?". Understanding human-object interactions is particularly important, as it can provide valuable insights into how humans interact with their environment.

The task of Human Object Interaction detection involves localizing a human and an object in an image, and identifying the interaction between them, which is defined by an action verb. Overall, the aim of HOI is to identify the triplet $< human, verb, object >$, which represents the interaction between a human and an object in a scene. This is an important step towards achieving high-level semantic understanding of the scene, as it allows us to infer the relationships between different elements in the image. As the field of HOI continues to evolve, we can expect to see many new and innovative approaches to detecting and understanding human-object interactions in images.

Conventional HOI methods can be divided into either two-stage methods or a one-stage methods. Most two-stage methods detect instances, and match the detected humans and objects one by one to form pair-wise proposals in the first stage. Next, in the second stage, such methods infer the interactions based on the features of cropped human-object pair-wise proposals. On the other hand, one-stage HOI detectors formulate HOI detection as a parallel detection problem, where interactions are localized with interaction points, or union boxes, replacing the separate neural network for interaction prediction with simple heuristic based matching methods which directly detect the HOI triplets from an image. One-stage methods have delivered great improvements in both efficiency and effectiveness.

With the recent introduction of transformers for contextual image embedding, one-stage methods improved the HOI detection by feeding the image deep features to a transformer encoder-decoder for output contextual embedding followed by a multilayer perceptron (MLP) for HOI triplet prediction. Therefore, one-stage methods can easily focus on the interactive human-object pairs and effectively extract corresponding features in an end-to-end manner.

When humans try to recall a painting, they rely on their visual memory and perform a search based on the elements present in the painting and the relationships between them. This is a complex process that involves identifying and recognizing the different objects and people in the painting, as well as understanding their interactions and the overall composition of the artwork. For people with little to no historical knowledge of art, this process can be particularly challenging. They may not be familiar with the style, genre, or period of the painting, which can make it difficult to interpret the visual cues and understand the intended meaning of the artwork.

Despite the importance of understanding human-object interactions in paintings, this area of research has not been thoroughly investigated, which presents a unique

opportunity for researchers to explore the potential of HOI in the interpretation of paintings and to develop new techniques for analyzing the relationships between different elements in the artwork. To achieve this goal, there is a need to create new datasets that can be used to train and test HOI algorithms in the context of paintings. These datasets must be carefully curated to include a diverse range of paintings from different periods and genres, and must be labeled with attributes such as the identity of the objects and people present in the painting, their interactions, and other relevant information.

2.3 Scene Understanding Work in Paintings

Various computer vision applications have been developed for scene understanding in paintings. Object detection and recognition in paintings are crucial for understanding the overall scene. Object recognition techniques are used to identify objects such as trees, buildings, animals, people, and other elements present in the painting. Researchers have also explored painting classification, which involves identifying the style and genre of a painting. This can help understand the artist's intentions and the context of the painting. Furthermore, image retrieval in paintings involves finding images that are similar to a given painting, and can be useful for art historians and curators who want to find paintings with similar themes, styles, or techniques. Several studies have been conducted in these areas, including works on deep learning-based object detection in paintings, paintings classification using deep neural networks, and content-based image retrieval in paintings.

2.3.1 *Object Detection in Paintings*

Object detection is used in paintings for the purpose of improving metadata to better support search, by making it possible to search for visual motifs and setting up new creative possibilities for artists and designers to create innovative art experiences.

In the study by Smirnov et al. [3] and Jeon et al. [4], the authors address the problem of the lack of labeled artwork datasets by generating new labeled artwork images through style transfer from a natural images dataset. Smirnov et al. [3] enhances object detection in digitized fine art by training data augmentation based on transferring styles from representative artworks to natural images. They fuse the style classification features with the object classification features with a SVM and show that including the style information increases the performance of the overall object classification. Jeon et al. [4] applies neural style transfer to the natural images in the COCO dataset for data augmentation and then trains an object detection method on the augmented dataset.

Moreover, Kadish et al. [5] addresses the cross depiction problem in paintings by using style transfer. Cross depiction refers to the tendency of the neural network to prioritize the identification of an object's texture over its shape. Style transfer is applied to the COCO dataset to build a model for only people detection in art images. The authors show the improvement achieved by their model in detecting people in paintings and drawings.

In addition, Marinescu et al. [6] improves object detection in paintings by combining Deep Learning and Semantic Metadata about candidate objects extracted from existing sources such as Wikidata, dictionaries, or Google NGram. The metadata refers to the time of first use of the words representing the objects and forms what they call a time matrix. The creation date of a painting is compared with the information in the timeholding structure to detect and replace anachronic objects with the most probable objects that fit the time period of the painting.

2.3.2 *Paintings Classification*

Paintings can be classified based on various attributes, including genre, style, artist, school, medium, and more. The genre of a painting refers to its category or type, which can include abstract, landscape, cityscape, religious, mythological, still life, and many others (Figure 2.2). On the other hand, painting style refers to the manner in which the art is expressed or performed, such as high renaissance, pointillism, realism, expressionism, minimalism, and others (Figure 2.3).

(a) Abstract painting. (b) Cityscape painting.

Figure 2.2: Different genres of paintings.

The process of labeling paintings based on these attributes can be performed either manually or automatically. Manual labeling requires individuals with an understanding of art history to analyze and assign labels to the paintings, which can be time-consuming and challenging, especially when dealing with a large collection of images. However, with advancements in computer science and machine learning, researchers have developed models and algorithms that can automatically label paintings based on different attributes.

Automatic labeling methods leverage techniques such as computer vision, pattern recognition, and machine learning to analyze the visual features of paintings and extract relevant information. These models are trained on labeled datasets, where human experts annotate paintings with their corresponding attributes. The models learn to recognize patterns and characteristics in the images, allowing them to automatically classify new paintings based on their genre, style, or other attributes.

Automatic labeling of paintings has the potential to significantly streamline the process of categorizing and organizing large art collections, making it easier for art historians, curators, and enthusiasts to search, browse, and study artworks. However, it is important to note that while automatic labeling can provide efficient

(a) High Renaissance painting. (b) Expressionism painting.

Figure 2.3: Different styles of paintings. Images from the MultitaskPainting100k dataset [7].

and consistent results, it may not always capture the nuanced interpretations and subjective aspects that human experts can provide. Therefore, a combination of manual and automatic labeling approaches can be a valuable approach to enhance the understanding and organization of paintings based on their attributes.

Before the emergence of Convolutional Neural Networks (CNNs), scientists relied on handcrafted features to represent input images in their models. Zujovic et al. [8] curated a dataset comprising 353 paintings from multiple websites like Artlex, Google, and CARLI Digital Collections, belonging to five distinct artistic genres: Abstract Impressionism, Cubism, Impressionism, Pop Art, and Realism. They utilized various feature extraction techniques such as Steerable Filter Decomposition for texture descriptors and edge detection to extract grey-scale features. Additionally, they extracted color features from the RGB images using the HSV (Hue, Saturation, Value) color model. The extracted features were then merged and fed to different classifiers such as AdaBoosted J48 decision tree, Naïve Bayes, K-NN, and SVM to classify the genre of the painting.

Zujovic et al. [8] created a dataset of 353 paintings from various websites, representing 5 genres: Abstract Impressionism, Cubism, Impressionism, Pop Art, and Realism. They used grayscale features like Steerable Filter Decomposition and edges, as well as color features extracted from the HSV values of RGB images. These features were used in classification tasks, where different classifiers such as AdaBoosted J48 decision tree [9], Naive Bayes, K-NN, and SVM were employed to identify the genre of each painting.

In their study, Culjak et al. [10] collected 693 images from Google and Artlex and classified them into six genres: realism, impressionism, cubism, fauvism, pointillism, and naive art. They used 68 features, including color and luminance histograms, edge detection, image sharpness, symmetry, and more. These features were fed to different classifiers such as ANN, RandomForest, SVM, k-NN, and decision table for

the final genre classification.

Agarwal et al. [11] developed separate models for style and genre classification using five features: SIFT [12], GIST [13], HoG [14] with LBP [15], GLCM [15], and color. They collected a dataset of paintings from Wikipaintings (http://wikipaintings.org) and selected six genres and ten styles for classification. Various classifiers, such as random forest, MLP, and libsvm with different kernels, were used for experimentation. The results showed that libsvm with X2 kernel provided the highest accuracy.

Lee et al. [16] developed a style classification model using 1633 paintings obtained from The Web Gallery of Art (http://www.wga.hu/index1.html). The system was tested on four painting styles: expressionism, impressionism, post-impressionism, and surrealism. The authors extracted 50 handcrafted color features, including average hue and average saturation, as well as composition features like shape and color by segment, to represent the global and local aspects of the paintings.

With the advancements in deep learning, Convolutional Neural Networks (CNNs) have become the standard for feature extraction in image recognition tasks. CNNs can automatically learn and extract relevant features from paintings, thereby enabling more accurate and efficient classification.

Saleh and Elgammal [17] developed a metric learning approach for predicting the style, genre, and artist of paintings. They optimized the similarity measures based on historical knowledge to project raw visual features into a new optimized feature space. Standard classifiers were then trained on this feature space for prediction purposes. The system utilized both classic visual descriptors and features learned by a CNN.

Cetinic et al. [18] conducted genre classification using a combination of six features: CNN-derived features, SIFT, GIST, HOG, GLCM, and HSV color histograms. They found that SVM was the most accurate classifier for genre classification. The experiment was performed on a subset of 1000 images from the WikiArt dataset, which included five different genres: history painting, religious painting, genre painting, landscape, and portrait.

Huang et al. [19] proposed a two-channel deep residual network for the classification of fine-art painting images. Their model incorporated both the RGB channel and brush stroke information channel. They pre-trained their model on ImageNet and utilized ResNet and AlexNet for feature extraction. Additionally, they employed the gray-level co-occurrence matrix to detect brush stroke information. The authors conducted separate experiments for genre, style, and artist classification using a dataset of paintings obtained from WikiArt.org. The dataset consisted of 25 styles, 10 genres, and 19 artists. They employed the SCM classifier as the final step in the classification process.

Hosain et al. [20] developed a deep learning model for feature extraction and classification of painting genres. They employed pretrained models, VGG-16 and Inception-V3, along with a modified CNN, for feature extraction. The final classification was performed using a softmax activation function. Their dataset comprised 3,215 images with 24 genres obtained from WikiArt. The experiments conducted by the authors demonstrated the ability of CNNs to improve the accuracy of genre

classification in paintings.

Wang et al. [21] introduced a Graph Neural Network (GNN) to model the potential relationship between the local styles of paintings. This enabled them to capture more discriminative and robust style information. Additionally, they designed a perceptual layer to learn cross-layer correlation features, which contributed to a stronger global style representation. For deep feature map extraction, VGG-19 was utilized. The authors used three style datasets from WikiArt: 2,338 images from Painting Styles, 4,266 images from Painting Genres, and 15,357 images from OilPainting. Their approach combined graphic style features with global style features to achieve improved style classification, leveraging the global consistency of the visual style.

Iliadis et al. [22] investigated the effectiveness of two different deep learning architectures, namely Vision Transformer and MLP Mixer, for artwork style recognition. These models were trained from scratch using the WikiArt paintings dataset, which consisted of 21 style classes. The performance of the two models was compared against popular pre-trained models like ResNet and VGG.

Menai et al. [23] employed EfficientNet, a pre-trained CNN on ImageNet, for feature extraction in the task of style classification for paintings. Transfer learning was applied to fine-tune different pre-trained EfficientNet models ranging from B0 to B6 on ImageNet to enhance classification accuracy. The experiments were conducted on the Painting-91 dataset, which comprised 4,266 digital paintings from 91 artists. The dataset included annotations for artist and style categorization tasks. The focus was on classifying 2,338 paintings from 50 painters into 13 art styles. The results demonstrated that deeper networks and higher input resolutions led to improved style recognition accuracy. Additionally, it was confirmed that deep retraining of layers, including the last fully connected layers, significantly contributed to enhancing the accuracy of style classification.

Zhao et al. [24] investigated the use of CNNs for art-related image classification tasks and proposed a big transfer learning (BiT) model. They demonstrated that models trained on real-world data could also be applied to the art domain. The authors examined how different hyperparameters affected model performance and found that higher resolution and appropriate training steps with mix-up improved accuracy. They systematically compared the performance of five weight initializations of the models for different tasks to evaluate the effectiveness of transfer learning. To validate their approach, three datasets were used for artist, genre, and style classification tasks. These included paintings with 14 styles and 91 artists from the Painting-91 dataset, paintings with 23 artists, 10 genres, and 27 styles from WikiArt, and paintings with 1,508 artists, 41 genres, and 125 styles from the MultitaskPainting100k dataset [7]. Additionally, an image retrieval system was built to enable users to find similar paintings based on artist, style, and genre features.

The focus of research in the field of art and paintings has been on understanding the artistic techniques, historical context, symbolism, and aesthetics of the artworks. There has been extensive research on various aspects of paintings, such as classification, style analysis, and art history, but the specific area of human-object interaction in paintings has received no attention. The study of human-object interaction in

paintings involves examining how human figures interact with objects or other elements within the artwork. This interaction provides insights into the narrative, cultural, social, or symbolic meanings embedded within the painting. By analyzing the relationships between humans and objects in paintings, researchers gain a deeper understanding of the intended message, storytelling, or cultural significance conveyed by the artist.

Chapter 3

Human Object Interaction

Human-Object Interaction (HOI) detection is a fundamental task for scene understanding, where beyond detecting individual humans or object instances in images, the interaction between them is also estimated. Accurate estimation of human-object interactions can benefit many downstream visual understanding tasks including image captioning, image retrieval, and visual question answering. HOI detection can provide valuable insights into the meaning and interpretation of a painting, as the interactions between humans and objects can reveal information about the scene, characters, and story depicted in the artwork.

3.1 Human Object Interaction

Given an input image, the aim of human-object interaction is to estimate and localize the interactions between humans and the objects around them by predicting the triplet $< human, predicate, object >$. Detecting these interactions requires both knowledge of human and object information, as well as the interactions between them.

The chronological order of HOI detection systems is presented in Figure 3.1, highlighting the increasing interest in this subject over time, which is largely attributed to advancements in deep learning methods and the availability of high-performance computers.

To tackle the HOI problem, there are different approaches that can be taken. One way is to detect all objects in an image and then consider every possible pairing between the detected human and objects. Different features are extracted from the detected parts to help predict the interaction between them. Another approach is to analyze the features extracted from the input image to locate the human and object, identify their class names, and determine the interaction between them. Recently, the introduction of vision transformers for object detection has significantly improved HOI detection. These transformer-based models use an encoder-decoder architecture to generate contextual representations of the interacting pair, along with a robust attention map for the interaction. As a result, HOI detection accuracy has been significantly enhanced.

Figure 3.1: Chronological order of HOI detection. Note the significant increase in work over the years due to the development of deep learning methods.

HOI can be broadly categorized as being either two stage or single stage. In two-stages approaches, the first stage is an object detection model and is used to localize the human and object bounding boxes along with their class labels; in the second stage, the detected pair is input to a neural network to extract features needed to predict the interaction between them. These features vary between visual, spatial, pose, and semantic features that are joined together differently, depending on the model, in order to predict the interaction label.

The second type of HOI detection methods are single staged, where features are extracted from the input image, and HOI triplets are directly detected from the image at the same time. Single stage methods specify the location of the interaction using an interaction point or an anchor box, then predict the action and the interacting pair bounding boxes along their class labels by matching the predictions with the detected objects. Later, with the introduction of transformers for contextual feature representation, single-stage methods were improved to end-to-end methods, where an image is processed by a CNN to extract 2D features that are input to a transformer encoder-decoder for human/object and HOI prediction and localization. Transformers rely on attention, which enables the models to focus more on certain parts of the input and thus reason more effectively. Using encoder-decoder attention over these embeddings, the model makes a judgement about all objects based on entire context of an image using pair-wise relations between them.

3.1.1 *Two-stage HOI Detection Methods*

The two-stage HOI detection methods (Figure 3.2) were first proposed in 2015. They first detect all possible interacting pairs, and then the human and objects cropped bounding boxes inputted into a CNN. From each image, features such as appearance, spatial, pose, and semantic features can be extracted from the pair, and then combined in different ways to predict the interaction between the candidate human/object pair. Most of the two-stage methods use off-the-shelf object detection models, and focus only on the architecture of the *interaction prediction* model.

Object Detection The first stage in a two-stage HOI detection system is object detection, which requires localizing an object inside an image and placing a rectangular bounding box around it, then associating to it a class label along with a prediction score.

A lot of improvement has been achieved in object detection in the past decade. Various methods rely on Faster RCNN pre-trained on the MS-COCO dataset [25].

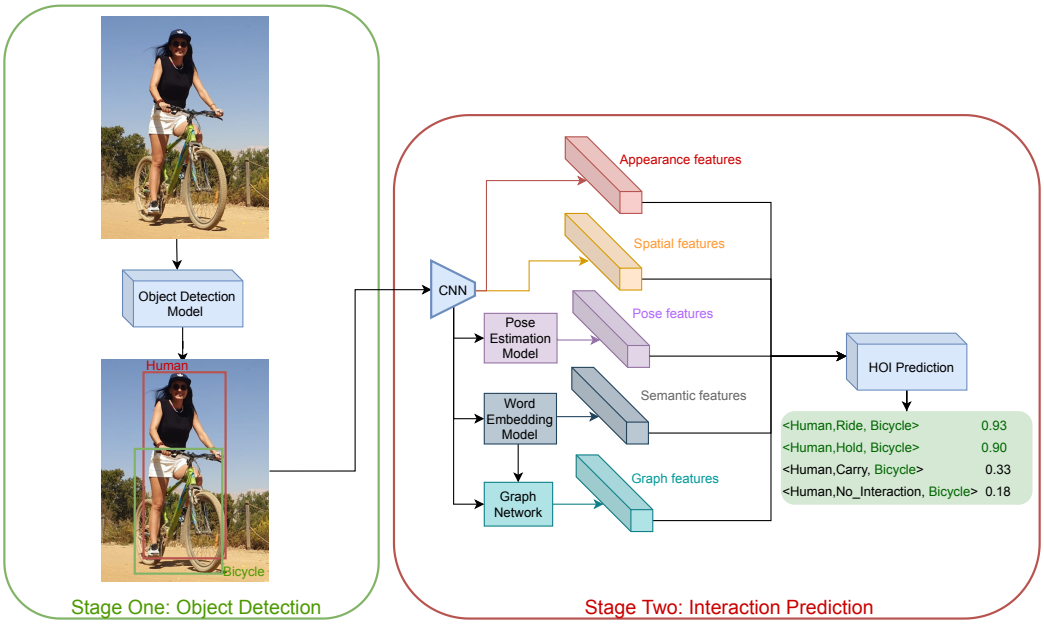


Figure 3.2: Two-stage Human Object Interaction Detection: in Stage One, all the humans and the objects are detected; in Stage Two, the interaction between the candidate human object pairs is predicted or classified. Here, human is interacting (riding, holding) with object (bicycle).

Other methods fine-tune the object detection model on one of the HOI datasets, such as HICO-DET [26], in an effort to obtain more accurate object detection predictions. After object detection is complete, objects with an output confidence score greater than a set threshold value are used for action classification. Typically, the threshold for the detected human is set between 0.5 and 0.8, and the threshold for other objects varies between 0.2 and 0.4 depending on the model. These thresholds are set to reduce the interaction prediction error caused by wrong object detections, thus filtering out wrong detections. The human threshold is set higher than that of the object to ensure that the model, while classifying the interaction, has the correct subject performing the action. Changing the score threshold allows tuning the false positive and true positive rates according to the specific needs of the model.

Interaction Prediction After identifying the interacting candidate human/object pairs, features are extracted and input to the interaction prediction network. Features include those related to appearance representations from the input image, spatial features from the detection bounding boxes, and pose features — where the human skeleton and facial expressions are detected. Moreover, semantic features can also be extracted, where object and action words are represented with their semantic representations using a word embedding model. The extracted features are fed either individually or jointly through a graph to the interaction prediction model to obtain the final action class representing the interaction between the candidate human-object pair. Table 3.1 lists the different features used by various two-stage HOI methods in the literature.

Table 3.1: Different visual features extracted by two-stage methods: appearance, spatial, pose. The appearance features can be extracted from the detected human, object, union, or from the entire input image.

Method	Appearance Features			
	Human	Object	Union	Entire image
HO-RCNN [26]	✓	✓		
InteractNet [27]	✓	✓		
GPNN [28]	✓	✓		
iCAN [29]	✓	✓		
Wang et al. [30]	✓	✓		✓
Bansal et al. [31]	✓	✓		
TIN [32]	✓	✓		
Xu et al. [33]	✓	✓		
Zhou et al. [34]	✓	✓		✓
PMFNet [35]	✓	✓	✓	
DRG [36]	✓	✓		
VCL [37]	✓	✓	✓	
VSGNet [38]	✓	✓		✓
DJ-RN [39]	✓	✓		
PD-Net [40]	✓	✓		
SCG [41]	✓	✓		
ConsNet [42]	✓	✓		
IDN [43]	✓	✓	✓	
VS-GATs [44]	✓	✓		
HOI-CL [45]	✓	✓	✓	

Appearance Features Appearance features account for the largest number of features that are used in the interaction model. Appearance features can be extracted, with deep networks, from the entire image, thereby providing context about the scene in which the interaction is happening. In addition, appearance features of each of the interacting pair are also extracted, those of the human alone yield information about the subject doing the action and, those of the object provide information about the object on which the action is happening. The object features are important in the HOI prediction because they represent the 'affordance' of the object which defines the possible uses of the objects in the real word. Moreover, features from the union of human-object bounding boxes represent the context in which the action is happening regardless of the surrounding. Note that appearance features from the human alone and object alone are always included (Table 3.1) in two-stage methods.

Appearance features are typically extracted with deep networks. ResNet (Residual Network) is one of the widely used feature extraction CNN for visual feature extraction. HOI models use different variants of ResNet to generate the visual feature maps from the input image. ResNet-50 [54] is one variant of ResNet consisting of 50 layers, ResNet50-FPN [55] is another variant of ResNet with an attached Fea-

Table 3.2: The CNNs used by the different models for visual feature extraction.

CNN	Year	Advantages	Disadvantages	Examples
CaffeNet	2014	Addresses overfitting and uses ReLU	Lower accuracy than other CNNs	[26]
ResNet-50	2016	Residual block solves the vanishing gradient problem	Less accurate than ResNet-101/152	[29], [30], [32]–[34], [37], [39], [43], [46], [47]
ResNet-101	2016	More accurate than ResNet-50	More training time and energy required	[28], [31], [37], [48]–[50]
ResNet152	2016	More accurate than ResNet-50/101	More training time and energy required	[38], [40], [51]
ResNet-50-FPN	2017	Constructs higher resolution layers from a semantic rich layer	No response to large-scale objects	[27], [35], [36], [40], [41], [44], [52], [53]
Deformable ConvNet	2017	Adapts to the geometric variations of objects	Features influenced by irrelevant image content	[28]

ture Pyramid Network used to construct a rich multi-scale feature pyramid from one single resolution input image. Moreover, ResNet101 and ResNet-152 are other variants of ResNet used in different HOI models, consisting of 101 and 152 layers respectively. In addition to ResNet, Deformable ConvNet [56] is another CNN used where both deformable convolution and RoI pooling modules have the same input and output as their plain versions. Table 3.2 lists the CNNs one can use to extract visual features from the input image along their advantages and disadvantages.

Spatial Features Spatial features are extracted from the detected human and object bounding boxes. It is difficult to estimate the distance between the interacting pair from a 2D image; instead, the relationship between the human and the object bounding boxes is used to encode the spatial relationship of the interacting pair. Spatial information can either be handcrafted or found using a feature map encoded by a CNN. Handcrafted features can include the Intersection over Union (IoU) of the bounding boxes, the distance between the boxes centers, or other relationships between the detected boxes as follows.

Xu et al. [33] use the normalized distance between the human and object bounding boxes and log of the ratio of the boxes widths and heights. Liang et al. [44] find the relative scale features and relative position features. The relative scale features include area ratio of each box to the image area and the ratio of each box top right

coordinate to the width and height. The relative position features include the ratio of boxes centers different to the image width and height, ratio of the difference of the top right corner coordinates to the width and height of the other box.

Liu et al. [42] construct the spatial configuration by finding the normalized distance between each coordinate of the detected bounding box and the union box origin. This distance is then normalized by the union box area. The spatial information used by [41] include center coordinates of the bounding boxes, widths, heights, aspect ratios and areas, all normalised by the corresponding dimension of the image. They also include the intersection over union, the area of the human box normalised by that of the object box, and another directional encoding. Bansal et al. [31] follow the work of [57] to define the geometric relationship feature in their model. Li et al. [39] construct 3D human-object spatial configuration by predicting the human 3D body from the 2D detection and finding the different relationships between the 3D human body parts and the object represented by a sphere.

Spatial feature maps are extracted between the bounding boxes using a CNN. First, a spatial attention feature map is created by adopting a two-channel binary image representation to characterize the interaction patterns. Second, the union of the two boxes is taken as the reference box, and a binary image is constructed with two channels within it. In the human channel, a value of 1 is assigned to the human bounding box and 0 elsewhere, while in the object channel, a value of 1 is assigned to the object bounding box and 0 elsewhere. The resulting two-channel tensor is then fed to a CNN to obtain spatial representations (as shown in Figure 3.3). Table 3.3 presents the two different methods to extract the spatial features with example models that adopt each of these methods.

Pose and Body-parts Stream In addition to the appearance and spatial features, some methods rely on the human pose as an additional feature to enrich their model (Table 3.4). Zhou et al. [34] use Detectron [59] of Mask R-CNN [60] trained on COCO training dataset for keypoint estimation including head, hand, hip and leg parts. Li et al. [32] and Zhong et al. [40] use RMPE [61] and CrowdPose [62] to estimate the detected human pose and add it to the spatial map to create a spatial-pose map. Wan et al. [35] use a Cascaded Pyramid Network (CPN) [63] as a pose estimator to estimate 17 keypoints of the detected human and use them

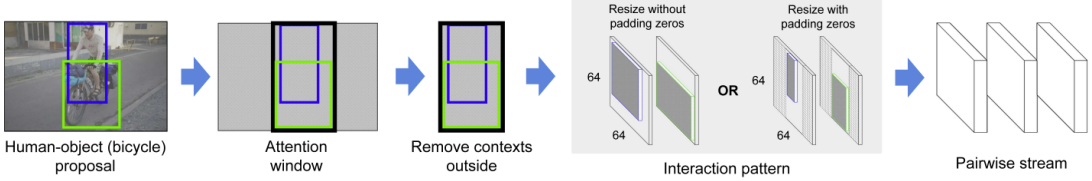


Figure 3.3: Construction of the binary maps (Interaction pattern) used to extract deep spatial features (Pairwise stream). Image from [58]

Table 3.3: Different spatial extraction methods used.

Spatial Extraction	Advantages	Disadvantages	Examples
Hand-crafted	Effective in encoding the spatial relationships	The choice of features is critical	[31], [33], [39], [41], [42], [44]
Binary-map	The network learns to find the correct features	Less efficient than hand-crafted	[29], [30], [32], [35]–[38], [40], [43], [45]

Table 3.4: Details of different pose and body part extraction models.

Pose/Body Part Extractor	Advantages	Disadvantages	Examples
Mask R-CNN	Estimates instance segmentation and keypoints together	Lower Precision	[34]
RMPE	Fast detection	Fails in crowded scenes	[32], [40]
CrowdPose	Efficient for crowded scenes	Depends on detection proposals	[32]
OpenPose	Accurate configuration	Slower detection	[39]
SMPLify-X	3D human estimation	Limited by a single 3D template mesh	[39]
CPN	Estimates occluded keypoints	Fails in crowded scenes	[35]

to capture subtle differences between similar interactions. Li et al. [39] first use OpenPose [64] to detect the 2D pose of body, face and hands. Then, they get the 3D human body using SMPLify-X and VPoser [65] by extracting the joint body, face and hands shape, face expression and pose which consists of jaw joints, finger joints, and body joints.

Semantic Stream Semantic word embedding in HOI is the process of representing the detected objects or actions in their semantic word representations. Language priors have been successfully used in many computer vision tasks, including visual relationship detection [66], image captioning, and visual question answering [67]. Word embedding features are considered semantic features because they generalize to the same object. Interactions are semantically related to each other; for example, a "person petting a horse" and a "person petting a zebra" are semantically similar. Both zebra and horse are animals and are represented close together in the word embedding space. Therefore, if the "person petting zebra" interaction is not seen frequently, it can be inferred from the "person petting horse" interaction. Therefore, semantically similar object can lead to similar interactions. Many two-stage models use semantic features in addition to the visual and spatial features to

Table 3.5: Details of the different two-stage models that used semantic features in their system. Ex.: Example, H: Human, O: Object, A: Action, BP: Body Parts.

Semantic Feature	Advantages	Disadvantages	Ex.	Semantic Features				
				H	O	A	H+A+O	BP
Extractor	Faster computation	No contextual embedding	[31]	✓	✓			
Word2vec			[39]		✓			✓
			[44]	✓	✓			
			[40]		✓	✓		
GloVe	More discrete vectors in the space	No contextual embedding	[33]		✓	✓		
FastText	Encodes unknown words	No contextual embedding	[36]		✓			
ELMo	Contextual embedding	Less accurate than BERT	[42]	✓	✓	✓	✓	
BERT	Contextual embedding	Needs GPU to run	[68]					✓

enrich their model and improve their HOI predictions. Table 3.5 presents the details of the semantic stream in different HOI prediction models.

One of the widely used word embedding models is Word2vec [69], which is a two-layer neural network trained to reconstruct linguistic contexts of words by leveraging the co-occurance within local context (neighbouring words). It is pre-trained on the Google News dataset and generates language prior features vector of dimension 600. Bansal et al. [31] cluster the objects based on their visual and semantic functional similarity using Word2vec and use these clusters to find all objects similar to an object in the target dataset. Using word embeddings in their model, they show that humans have similar interactions with objects that are functionally similar. Word2vec is also used by Li et al. [39] to pair the extracted spatial features of the detected object and human body parts with their PCA reduced word embedding features. Moreover, Zhong et al. [40] use Word2vec to generate the word embeddings of the object and verb categories and show that the word embedding of the verb and the object together as an input to the model perform better than only using the verb alone, that indicates the importance of the presence of the object word embedding in the language prior to solve the verb polysemy problem which is the coexistence of many possible meanings for a word or phrase.

Other HOI systems rely on GloVe (Global Vectors for Word Representation) text model [70], trained on the Wikipedia dataset, to generate their semantic features which is an unsupervised learning algorithm trained on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space. For example, Xu et al. [33] learn semantic structure aware embedding space compared to original word embeddings where they build a graph that can leverage semantic similarity to retrieve the verb best describing the detected human and object pair using the GloVe model.

GloVe captures global contextual information in a text corpus by calculating a global word-word co-occurrence matrix. Whereas, Word2Vec only captures the local context of words. During training, it only considers neighboring words to capture the context while GloVe considers the entire corpus and creates a large matrix that can capture the co-occurrence of words within the corpus.

FastText [71], developed by Facebook released in 2016, follows the same idea as Word2vec but with a major twist. Instead of using words to build word embeddings, fastText uses parts of words and characters, where a word becomes its context. The building stones are therefore characters instead of words. The word embeddings outputted by fastText look very similar to the ones provided by Word2Vec. However, they are not calculated directly. Instead, they are a combination of lower-level embeddings. Gao et al. [36] use fastText to generate a vector representing the object category and show that the addition of semantic features to the spatial features improves the HOI prediction because it enables knowledge transfer between object classes and helps with rare interaction during training and inference.

FastText solves one of the main disadvantages of Word2Vec and GloVe embedding, which is the encoding of unknown or out-of-vocabulary words that can be represented in vector form as it has high probability that its n-grams are also present in other words.

ELMo (Embeddings from Language Models) [72] is a deep contextualized word representation that models both complex characteristics of word use and how these uses vary across linguistic contexts. Word vectors are learned functions of the internal states of a deep bidirectional language model (biLM), which is pre-trained on a corpus of 5.5 billion words. Liu et al. [42] use the word embeddings of the interaction triplets as input to features of the nodes in their graph attention network in order to learn implicit knowledge of HOIs. They show how their model improved by changing word embeddings from Word2vec [69], GloVe [70], or fastText [73] to ELMo [72] and show that ELMo performed better than the other word embedding models because it captures contextual information and considers the triplets jointly as a whole.

Unlike traditional word embeddings such as Word2vec and GLoVe, the ELMo vector assigned to a token or word is actually a function of the entire sentence containing that word. Therefore, the same word can have different word vectors under different contexts.

BERT (Bidirectional Encoder Representations from Transformers) [74], published by Google, is a language understanding model that considers the context of words and uses a deep bidirectional transformer to extract contextual representations by generating a language prior vector of size 1x768. BERT is pre-trained the whole of the English Wikipedia and Brown Corpus.

Li et al. [68] use BERT to map the Human Body Part States (PaSta) found from the pose, to the activity semantics of each body part with respect to the detected object. In their studies, they show that using BERT as a word embedding model performed better than Word2Vec, GloVe, and Gaussian noise which is due to the fact that BERT gives contextual information about the input words from the sentences and enriching its representation.

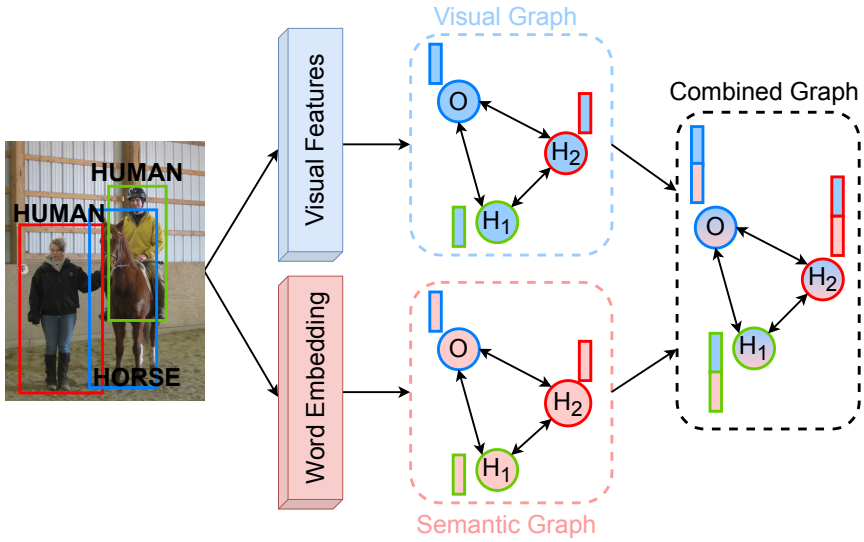


Figure 3.4: Example graph networks constructed from the detected humans and objects in the image. In the visual graph, the nodes are represented by their visual features. In the semantic graph, the nodes are represented by their word embeddings. In the combined graph, the nodes are represented by the concatenation of the visual and semantic features. Image from HICO-DET dataset [26].

Graph Neural Network Graph Neural Networks (GNNs) have been used to model scene relations and knowledge structures through a graph. A graph $\mathcal{G}$ is composed of nodes N that are connected together with edges E. A GNN learns potential features for nodes by iteratively propagating messages from neighboring nodes and updating hidden states using embedded functions. Different types of graphs can be used for better HOI predictions: visual graphs, semantic graphs, and combined visual and semantic graphs (Figure 3.4). In visual graphs, the nodes are represented by appearance features, while in semantic graphs, the nodes are represented by their word embeddings. In combined graphs, the output embeddings from the visual and semantic graphs are concatenated together to represent the node. By improving the performance of HOI detection models, graph networks have shown the ability to model high-level structures and leverage the learning capabilities of the network. Table 3.6 presented the graph-related design decisions of the models that used graph network in their system.

Visual Graph: Visual graph neural networks relate the appearance features of the detected humans and objects from the input image. The edges connecting the nodes can be represented by the spatial features generated from the detection bounding boxes such as in [28], [36], [41], [75]. Qi et al. [28] extract the edge feature from a combined bounding box, that is, the smallest bounding box that covers the human and the object together. Zhang et al. [41] build a spatially conditioned graph; they construct a bipartite graph where the edge features are computed as handcrafted feature vectors using spatial information from the centre coordinates of the bounding boxes, widths, heights, aspect ratios and areas. They also include the intersection over union, the area of the human box normalised by that of the object

box, and a directional encoding using the normalized differences between centre coordinates of the human and object boxes. They demonstrate the advantages of spatial conditioning for the computation of the adjacency structure, messages and the refined graph features.

Wang et al. [75] argued that the graph should take into consideration the fact that there are two sets of heterogeneous nodes: humans nodes and objects nodes. Thus, message passing between homogeneous nodes (intra-class messages) are modelled separately from that between heterogeneous nodes (inter-class messages). The spatial relation of a person and an object constitutes essential information for recognizing the interaction and is encoded into the edges that connect heterogeneous nodes. Gao et al. [36] also took advantage of the heterogeneity in nodes by constructing separate human-centric and object-centric graphs. They modelled human–object pairs as nodes, and employed the spatial-semantic features as node encodings where the spatial features are extracted using the two-channel binary image representation and the the semantic features represent the word embedding of each object's category, using fastText.

Another way to represent the edges in a visual graph is by using the interaction proposal scores such as in [38]. Ulutan et al. [38] represent the edges by the interaction proposal scores that are generated from the spatially refined visual features.

These models have proven that the addition of the graph branch to the visual-spatial branch improved the model's performance. Nonetheless, using a good object detector with accurate bounding boxes and finding the best edge representation is essential in increasing the performance of the visual graph network.

Semantic Graph: In the semantic graph network, the nodes are represented by their word embeddings and can encode the objects, action verbs, the interaction sentence, or even the human body parts. The graph edges can encode the spatial relationships, the statistical co-occurrance, or other features.

For example, Xu et al. [33] construct a knowledge graph where the nodes of the are represented by the Word2vec embedding of each object and action word and the edges are represented by binary values defining the connection or the disconnection of the nodes based on the ground-truth annotations of training dataset and external source. Only the candidate verb features from the updated nodes are extracted and compared to the visual features of the human minus object for joint embedding learning. Their results indicate that modeling semantic dependencies of verbs-objects in relationships and leveraging message passing capabilities of Graph Convolutional Networks (GCNs) together is essential for HOI prediction.

Liu et al. [42] encode the relations among objects, actions and interactions into an undirected graph called consistency graph including an HOI represented by four nodes: object node, human node, action node, and one node for the interaction that includes the triplet in a sentence. These nodes are represented by their semantic features obtained from ELMO [72] and the the edges are defined based on the consistencies among objects, actions, and interactions. If two nodes are semantically consistent with each other, an edge would be added to enable message passing between them. They show the significance of their proposed knowledge-aware strategy by comparing their model with and without using semantic embeddings.

Table 3.6: Details of the different two-stage model that used a graph network in their system. Ex.: Example, H: Human, O: Object, A: Action, I: Interaction

Graph	Advantages	Disadvantages	Ex.	Nodes	Edges
Visual	Captures contextual cues	Confuses model by plausible spatial configuration	[28]	H & O	Bounding Box
			[36]	H & O	Spatial-Semantic Features
			[75]	H & O	Spatial Features
			[38]	H & O	Interaction proposal scores
			[41]	H & O	Spatial Features
Semantic	Aids detection despite diverse scenes	Does not aggregate visual and spatial cues	[33]	O & A	Binary values for connection
			[42]	H & O & A & I	Consistency-based edge
			[68]	Body parts	Statistical co-occurrence
Combined	Richer representation of instances	Visual and semantic cues may have orthogonality to them	[44]	H & O Visual	Spatial Features
				H & O Semantic	Spatial Features

Li et al. [68] construct a Hierarchical Activity Graph (HAG) to model the activities for HOI prediction. Based on human part-level semantics, they built a large-scale part state knowledge base and Activity2Vec for finer-grained action encoding where the node represents the word embedding of the different human body parts and the edge indicates the statistical co-occurrence.

Semantic-based graphs have proven to improve the performance of the HOI detection models especially on the rare interactions of the HOI-DET dataset. This is due to the fact that the graph neural network is responsible for transferring the knowledge from non-rare to rare classes.

Combined Visual and Semantic Graph: Some methods combine both visual and semantic graphs in their systems to improve the HOI prediction. By combining the visual and the semantic cues, the model generates richer representations.

Liang et al. [44] build a dual-graph attention network that aggregates visual, spatial, and semantic information of the detected human-object pairs. The first built graph is a visual graph, where the nodes represent the detected human and object's appearance features and the edges connecting the nodes are represented by the spatial features relating the object nodes. The second graph is a semantic graph, where the human and object nodes are represented by their word embedding (Word2vec) features and spatial features are used to instantiate the edges in the final combined graph. After inputting each of the graphs in a Graph Attention Network (GAT), the node features are updated by aggregating its neighboring node's features and the resulting node features from the visual and semantic graphs are concatenated

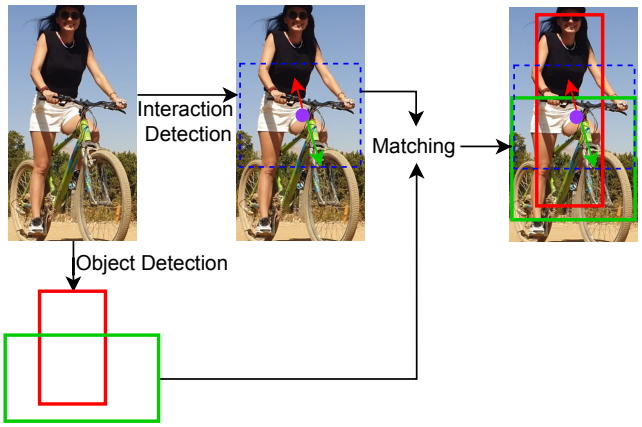


Figure 3.5: One-Stage HOI detection: Object and interaction detection are preformed synchronically . Then, the interaction class is matched with the detected instances for output interaction triplet prediction.

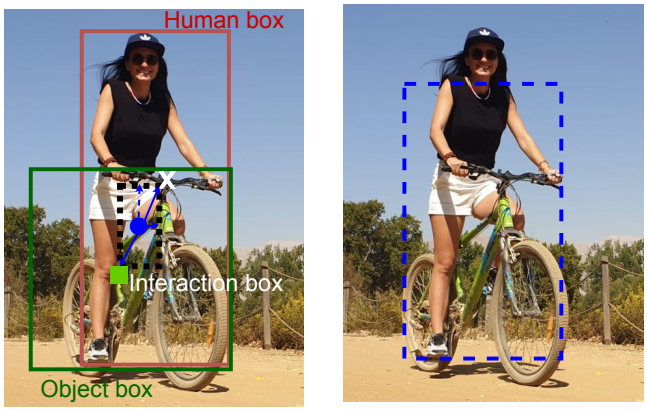
together to represent the human and objects nodes in the combined graph. Their results show that semantic cues can promote HOI detection but less effective than visual cues.

Graph neural networks have been used by HOI methods to model the relationship between humans, their body parts, the objects and the actions in order to help predict the correct interaction for each pair. Through GNNs, contextual information is propagated between nodes which enhances the HOI prediction compared to interpreting each detected instance in isolation or just relying on their 2D spatial relationship.

3.1.2 *One-stage HOI Methods*

In one-stage HOI, the interacting pairs and the interactions are detected at the same time, and the interaction triplet is estimated using two parallel branches: in the first branch, the image feature map is used to find the interaction point or the interaction area and accordingly the interaction. In the second branch, object detection is performed to find the human-object pair. Finally, these two branches are used to regress the offsets and match the interaction class with the detected instances (Figure 3.5).

Single-stage methods have shown to be more efficient than two-stage methods that require two separate and unrelated steps and limit their performance based on the quality of proposals in the first stage. Moreover, two-stage methods miss contextual features from the images even when the union of the human and object is taken into account. It is true that features from the bounding box joining the human and the object of interest contain context, but they also contain disturbing information that can come from the background or other objects present in that box. Single-stage methods capture contextual information from the image by pairing the target human and object from an early stage in feature extraction and extracting integrated features rather than individually treating the targets.



(a) (b)

Figure 3.6: One-stage methods: (a) Point-based method: the cross indicates the human point, the square indicated the object point and the circle indicated the interaction point. (b) Anchor-based method: the dashed box indicates the interaction box.

One-stage methods can be categorized as either point-based (Figure 3.6a) or anchor-based (Figure 3.6b). Table 3.7 presents the different models that detected the human object interaction using one-stage methods. Point-based methods perform inference at each interaction key point, such as the midpoint of each corresponding human-object pair. The key component in point-based methods is the interaction point detection where the extracted appearance features, using Hourglass-104 [76] are used to estimate the center point of each of the human and object with their corresponding sizes (width and height), in addition to the local offsets. This detected interaction point provides context and regularization to the human and object detection boxes and is used in the point matching branch, where the displacements from the interaction point are first regressed to the human point and object point respectively. Then, the interacting triplets are generated by matching each interaction point with the human and object points.

Point-based methods were adopted by Liao et al. [77] and Wang et al. [78] where they generate an interaction point from the visual features extracted from the image. At the same time, an object detector is applied on the image and finally used with the interaction point for final triplets generation in the interaction matching branch.

GGNet by [79] is another point-based method which predicts interactions more robustly using two steps: first, every pixel is interpreted in the input image feature map and identified whether it is a possible interaction point or not. In the second step, they interpret the feature map generated from the first step with possible interaction points to finally find a refined set of ActPoints. The features of the refined ActPoints are finally used to infer the interaction at at the interaction points. In addition to the interaction points identification, the interaction human-object pair matching method is enhanced by assigning to each interaction category a unique location regressor. This way, the effect of the interaction category on the spatial layout of one human-object pair is reduced.

Table 3.7: Advantages and disadvantages of the different one-stage HOI detection methods.

One-stage Methods	Advantages	Disadvantages	Examples
Point-based	Directly detects interactions between pairs as a set of interaction points	No apparent characteristics in visual patterns	[77]–[79]
Anchor-based	Directly detects the interaction region	Not straight-forward and sensitive to occlusion	[53], [80]

Although interaction points converge the HOI instance detection and recognition together, there are mainly two drawbacks; first, the semantic features are ambiguous when the interaction point is far apart from the human and object; second, the lack of a multi-scale architecture which is commonly used in object detection. On the other hand, anchor-based methods predict the interactions based on an anchor box instead of a interaction point.

Kim et al. [53] propose an anchor-based HOI detection method called UnionDet where the union bounding box of a human-object pair is detected and used to extract integrated features. They use a union-level detection framework to directly capture the region of interaction and an instance-level detector to perform object detection and action classification. To allow for more accurate instance-level localization, they combine the union-level detector and an instance-level detector in parallel.

Another anchor-based HOI detection method is built by Fang et al. [80], where they propose a Dense Interaction Region Voting (DIRV) framework that concentrates on the interaction regions of the human-object pairs. Their interaction region denotes the region that covers the minimal area of human and object crucial for recognizing the interaction and does not need to cover the whole human and object. To specify the interaction region, they set thresholds on the human bounding box, object bounding box, anchor box, and the union region.

Through one-stage methods, HOI detection performance was improved because contextual information was captured from the image by pairing the target human and object from an early stage in feature extraction, in addition to the accelerated training and testing times due to the fact that both object detection and HOI prediction were done in parallel and not separately. However, one-stage methods are limited by complex handcrafted grouping strategies to group object detection results and the interaction predictions into final HOI triplets.

3.1.3 *End-to-End Transformer based methods*

Transformers have had success with Natural Language Processing (NLP) and recently applied to images with image transformers. While CNNs use pixel arrays, self-attention mechanisms of transformers explicitly model all pairwise interactions between elements in a sequence, making these architectures particularly suitable

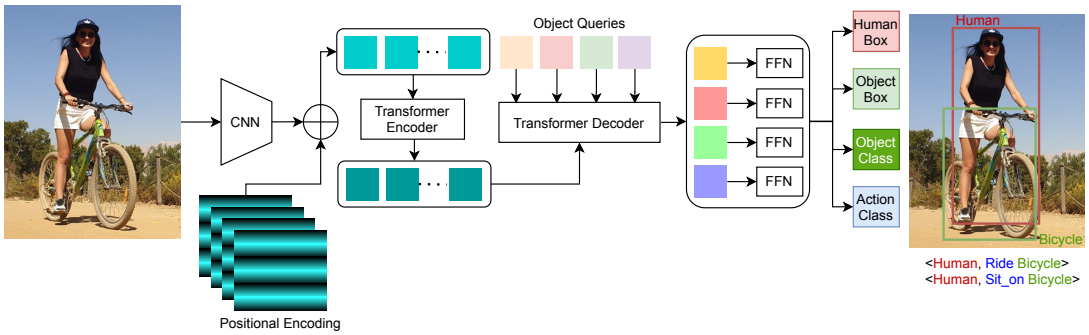


Figure 3.7: End-to-end transformer-based HOI detection: The image is fed to a CNN backbone to extract visual feature. Then, the feature is reduced in channel-dimension, flatten in spatial-dimension and complemented by positional encoding. The transformer encoder generates the global memory feature and the transformer decoder transforms the learnt positional embeddings into output embeddings. Finally, multiple Feed Forward Networks (FFN) predict the four HOI instances simultaneously.

for specific constraints of set prediction such as removing duplicate predictions. Single-stage methods require a trade-off between the interaction classification and human-object pair positioning, which is time expensive too. For this reason, end-to-end transformer based HOI methods were proposed by [48], [49], [46], [47], and [50]. In transformer-based methods,the model detects the human-object pair in addition to the interaction class in one shot, without any post-processing (Figure 3.7). One key component in these methods is the self-attention mechanisms of transformers which makes the model exploit the contextual relationships between human and object and their interactions to predict better the set of HOI triplets.

Transformer-based models consist of three parts: In the first part, the image is fed to a network for visual feature extraction. Many models rely on Detection Transformer (DETR) [81], which is an object detection system based on transformers. It is built by combining a CNN, such as ResNet-50 or ResNet-101, with a transformer encoder for visual feature representation.

In the second part, the extracted feature map is inputted to a transformer decoder to produce the output embeddings. The transformer decoder transforms a set of learnable query vectors into embeddings that contain image-wide contextual information for HOI detection, referring to the encoded feature map using the attention mechanism. The attention mechanisms in the transformer decoder are the key components that model the relations between feature representations of different detections in their systems. Zou et al. [48] and Tamura et al. [49] use a single decoder for instance and interaction representation, whereas Kim et al. [46] and Chen et al. [47] propose a transformer-based two-branch architecture. This architecture constructs an instance decoder and an interaction decoder to decode the boxes and action classes of the HOI instances in parallel. The instance decoder transforms the instance queries into instance representations for object detection, while the interaction decoder transforms the interaction queries into interaction representations for

interaction detection. By using two separate decoders for instance and interaction representation, the model is able to predict the HOI triplets more accurately by learning effective features through separate focusing on detecting human and object pairs and defining the interaction locations. Chen et al. [47] map a trainable interaction query set to an interaction prediction set with a transformer and design an effective instance-aware attention module to introduce instructive features from the instance branch into the interaction branch.

In the third part, the generated decoder representations are fed to several multilayer perceptron (MLP) layers to generate the final HOI quintuples, which consist of the human box, object box, object class, and action class. Zou et al. [48] and Tamura et al. [49] use three one-layer MLP branches to predict the human confidence, object confidence, and interaction confidence respectively, and two three-layer MLP branches to predict the human box and object box. Kim et al. [46] use feed-forward networks (FFNs) for the interaction representation to obtain a Human Pointer, an Object Pointer, and interaction type. They localize the human and object regions by pointing to the relevant instance representations using the Human Pointer and Object Pointer (HO Pointers), instead of directly regressing the bounding box. Then, they apply the feed-forward networks for bounding box regression and action classification to generate the final HOI quintuples. Chen et al. [47] use an FFN head on top of each decoder layer to decode a set of instance and interaction predictions. The instance FFN head comprises three independent sub-branches: one to predict the normalized bounding box for each detected instance, another to infer scores for object categories, and the last to generate a distinctive semantic embedding. The interaction FFN head is also split into three sub-branches: a 4-dimensional interaction vector with categories, and two semantic embeddings for the corresponding human and object instances, respectively.

Zhou et al. [82] decouple the triplet prediction into human-object pair detection and interaction classification via an instance encoder-decoder stream and an interaction encoder-decoder stream, where both the encoder and decoder are disentangled. By disentangling the encoder, they generate relations in image representations designed for each sub-task. To create communication between the decoders, they use an attentional fusion block by fusing the instance representation to the interaction representation as if they are associated with the same query index.

Chen et al. [83] improved transformer-based methods by using query-based anchors to extract HOI embeddings and predict HOI instances. Instead of relying on a CNN network and DETR, they extracted multi-scale features by combining a hierarchical network and a deformable DETR encoder with Swin-Transformer [84]. The transformer decoder and the interaction head used query-based anchors to decode the HOI embeddings and predict the HOI instances. Through their experiments, they demonstrated the effect of multi-scale feature maps and transformer-based networks in improving prediction accuracies by capturing non-local semantic features and spatial information.

Dong et al. [85] promoted the HOI detector by initializing the object query with category-aware semantic information instead of zeros, as done in previous methods.

Table 3.8: Design Decision of Transformer-based HOI detection methods. Trans.: Transformer, Enc:Encoder, Dec:Decoder

Trans. Structure	Design Decision	Advantages	Disadvantages	Example
Positional Encoder	Learnable embedding	Updated with network weights	Unable to adapt to longer input sequences during testing	[47], [48], [85]
	Sinusoidal embedding	Relative positioning	Changes with translation	[46], [49], [50], [83]
Encoder	6-layer Enc	Contextualized representation of image	Cannot encode images with small objects	[46]–[50], [85]
	6-layer Instance Enc and 6-layer Interaction Enc	Learns representations for each sub-task	High Complexity	[82]
	6-layer deformable Enc	Multi-scale semantic feature extraction	Lack of keys restricts representation power	[83]
Decoder	6-layer Dec	Self-attention mechanism	Cannot detect HOI instances with small objects	[48], [49], [85]
	6-layer Instance Dec and 6-layer Interaction Dec	Attention on each sub-task	Cannot detect HOI instances with small objects	[46], [47], [50], [82]
	6-layer deformable Dec	Multi-scale deformable attention	Requires multi-scale feature maps for accurate detection	[83]

Table 3.8 presents the design decisions of transformer-based HOI detection methods. Transformer-based models showcased the importance of transformers, especially for feature representation. The attention maps generated by transformers were able to capture contextual information about the interaction. They were able to aggregate only the relevant information without the need to crop the image and without being affected by unrelated instances present between the interacting pair, which could contaminate the interaction representation.

3.1.4 *Related Work on HOI Detection*

Researchers have solved the HOI problem using two types of methods: two-stage and single-stage methods. The first stage in a two-stage method is the detection of the humans and objects using an off-the-shelf detector, then in the second stage the interaction between them is predicted using the extracted features. In single-stage systems, the object detection and interaction prediction are done in parallel or in an end-to-end manner. Most of the existing two-stage systems, [27], [29], [37], [43], [86], rely on interpreting the scene based on its appearance as well as the geometric layout of objects and people within the scene. In some of these works, contextual information is only incorporated through features from the union region of a human and object bounding box which may not always be shown in the features covering the union region. Other systems, [32], [35], [87] solve the HOI problem by estimating the pose of detected people as an addition to the spatial and visual features. Other two-stage networks, [31], [33], [36], [42], [44], [68], [88]–[92], predict the HOI prediction by integrating semantics into the network architecture. Xu [33] construct a knowledge graph between object and action based on the semantic features of the ground-truth annotations of training dataset and external source. Bansal [31] integrate visual and spatial features with general word embedding of humans and objects. Gao [36] propose a dual relation graph by using spatial-semantic representation to describe each human-object pair. Liu [42] build a consistency graph that encodes the relations among objects, actions and interactions. Liang [44] build a dual-graph attention network that aggregates contextual visual, spatial, and semantic information.

To improve the HOI detection, recent works have developed one-stage pipelines to detect HOIs in a single shot. Single-stage methods, [53], [77], [78], localize the interaction with an interaction point or find the anchor box of a human-object pair. Contextual features are extracted around the detected point or box. The interacting triplets are predicted by matching the detected objects with the localized interaction and manually searching for the threshold. Later, single-stage methods were improved by using end-to-end transformer-based methods [46]–[49]. A transformer-based contextual self-attention mechanism is used to detect the interacting pairs and predict their interaction simultaneously. In these single-stage methods, contextual features are extracted visually from the image without any semantic representations. However, relying on visual context can be tricky in images where details are not well visible, such as in paintings and artwork.

Chen et al. [83] improve the transformer-based methods by using query-based anchors to extract the HOI embeddings and predict the HOI instances. Instead of relying on a CNN network and DETR, they extract multi-scale features by combining a hierarchical network and a deformable DETR encoder, with Swin-Transformer [84]. The transformer decoder and the interaction head use the query-based anchors to decode the HOI embeddings and predict the HOI instances. Through their experiments, they showed the effect of multi-scale features maps and transformer-based network in improving the prediction accuracies by capturing non-local semantic features and spatial information.

The field of human-object interaction detection has been a popular area of re-

search in computer vision in recent years, with many state-of-the-art models and techniques being developed and tested on natural photographic images. However, there has been very little research conducted on the detection of HOIs in paintings, which represents a unique challenge due to the artistic nature of these images.

Detecting HOIs in paintings poses several challenges, such as the highly stylized and often unrealistic depiction of humans and objects, as well as the lack of contextual information that is present in photographic images. In addition, paintings often contain many visual elements that are unrelated to the HOI of interest, which can make it difficult to accurately detect and classify interactions.

Given these challenges, there is a significant need for research to explore the detection of HOIs in paintings. By developing and testing new models and techniques that are specifically designed to handle the unique characteristics of paintings, researchers can gain a better understanding of how to analyze and interpret this type of artwork. This could lead to new insights into the cultural and historical context of paintings, as well as new applications in areas such as art conservation and museum curation. To the best of our knowledge, no previous research has been conducted on the detection of human-object interactions in paintings.

3.2 Difference between Natural Images and Paintings for HOI

In natural images, the context plays a crucial role in the detection of human-object interactions. However, in paintings, the concept of context may not always be applicable since artists can intentionally violate reality to convey their message or express their creativity. Therefore, detecting human-object interactions in paintings can be challenging since common sense may not hold true in such scenarios. For instance, physical rules such as gravity, scale, lighting, shading, and color may not be bound by reality in paintings, making it difficult to apply traditional HOI approaches.

Moreover, paintings are not always realistic, and the artist's subjective view of the world may significantly impact the representation of objects and humans in the painting. This can result in the creation of objects that do not exist in reality or the deformation of humans to convey specific emotions or feelings, making the detection of human-object interactions in paintings more complex. Therefore, it requires a different approach that is tailored to the unique characteristics of paintings.

Furthermore, paintings can be heavily influenced by the historical, social, and cultural contexts in which they were created, adding an additional layer of complexity to the interpretation of human-object interactions. Artists may use specific symbols or metaphors that were prevalent during a particular time period or cultural movement, and detecting these interactions may require knowledge of the broader historical and cultural context.

Additionally, the interpretation of human-object interactions in paintings may also depend on the viewer's subjective understanding and interpretation. Different viewers may have varying levels of knowledge and experiences, leading to different interpretations of the same painting. Thus, the detection of human-object interactions

in paintings may require a more nuanced and subjective approach than traditional HOI detection in natural images.

In the context of HOI detection in paintings, human pose features may not be as valuable compared to other types of features. This is because paintings often depict humans in stylized or exaggerated poses that may not correspond directly to realistic human poses. The artistic interpretation and style of the painting may prioritize aesthetic or symbolic representation over anatomical accuracy.

Overall, the challenges of detecting human-object interactions in paintings are multifaceted and require a unique approach that considers the complex interplay between the artistic representation, historical and cultural context, and the subjective interpretation of the viewer.

Chapter 4

Proposed HOI System in Paintings

In this chapter, I introduce the proposed architectures for HOI detection models in paintings. Two different models are presented: a single-task learning model, where the detection of interacting pairs and their corresponding interaction verb prediction are performed jointly, and a multi-task learning model, where HOI detection is improved by detecting it along with the classification of four other painting-related tasks.

4.1 System Overview

The visual search for targets in a scene is guided by the interplay between two types of memory: episodic memory and semantic memory. Episodic memory, housed in the hippocampus, is responsible for encoding and recalling specific details such as the position, colors, edges, and contextual information associated with the object from previous experiences. On the other hand, semantic memory, located in the neocortex, encompasses general knowledge, facts, concepts, and ideas that are independent of personal experiences. Semantic memory provides answers to more general questions about objects, including their affordance, name, type, or typical position. These two types of memory are not isolated but interconnected. Semantic memory is built upon the accumulation and integration of episodic memories. The episodic memories contribute specific details and contextual information, while the semantic memory extracts the general information and concepts that emerge from these episodic experiences. Together, these memories work in concert to form a comprehensive understanding and representation of the scene [93].

The proposed networks are inspired by human psychology, which recognizes that the human brain relies on both the actor and the object to infer an interaction. Similarly, in the proposed networks, I leverage this understanding by considering not only the human performing the action but also the object they are interacting with. The proposed networks aim to bridge the gap between visual and semantic information, connecting the representation of the visual scene with the contextual understanding of the action and object in paintings. This approach aligns with the

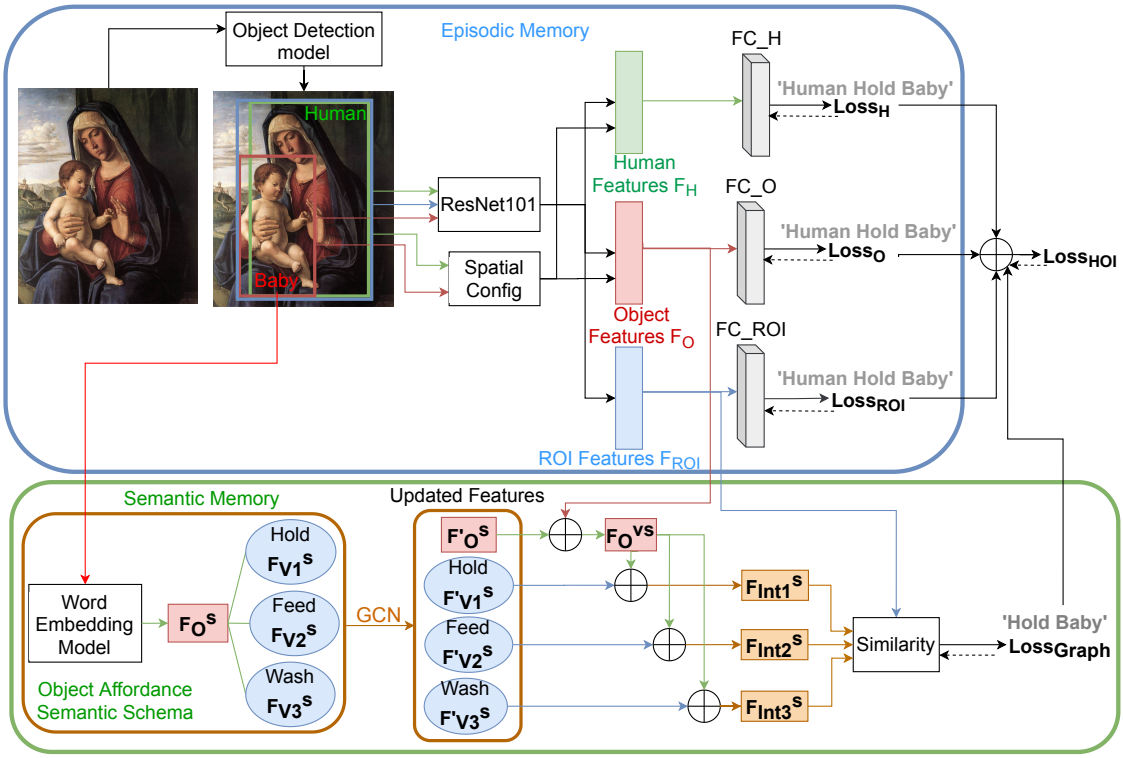


Figure 4.1: System flowchart: It consists of an episodic memory module where visual features (Figure 4.2) and spatial features (Figure 4.3) are extracted for the object, human, and the region containing them. both visual and spatial features are concatenated together to generate the human and object appearance features (Figure 4.4). In the semantic memory module, semantic features are extracted for the detected object and all the related candidate actions. The input features of the detected object-related actions are replaced with their contextual ones. Losses are calculated from the episodic and semantic modules and joint together for optimization.

interplay between visual and semantic memories observed in human psychology, as it combines both modalities to enhance the inference of human-object interactions.

4.1.1 *Single-task learning on paintings*

Figure 4.1 illustrates the flowchart of the proposed two-stage HOI detection system, **HOI-Paint**. The primary goal of this system is to predict the triplet *<human, verb, object>* for each candidate human-object pair within the input image. From a psychological perspective, our visual system initially focuses on detecting and localizing objects within the scene [94]. This process involves identifying unique shapes, colors, and textures that correspond to different objects, forming the foundation of our mental representation of the visual scene. Hence, the first stage of the proposed system involves detecting all the candidate instances present in the image.

After object detection, human visual perception abstracts relevant features from the raw sensory input [95]. Similar to how humans analyze objects by detecting dis-

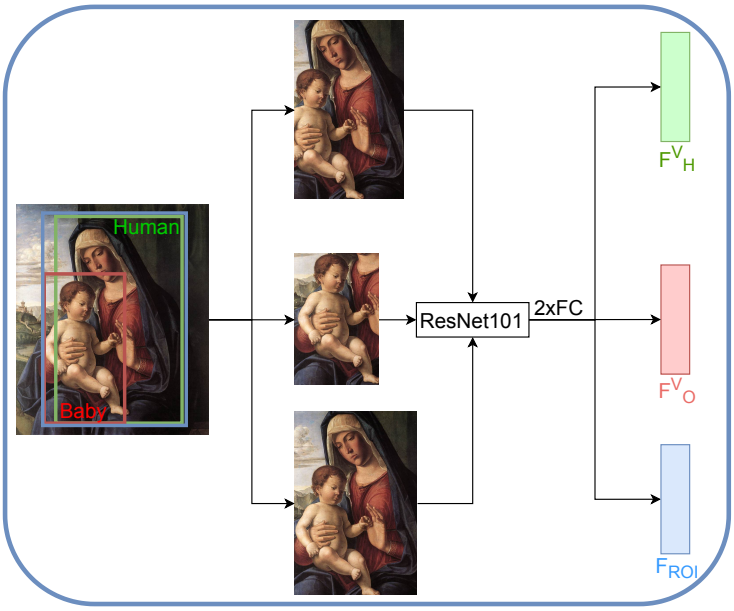


Figure 4.2: Visual stream of the proposed system.

criminative features, the proposed model learns to identify and represent distinctive visual patterns enabling it to capture essential discriminative information. Therefore, visual features are extracted from the detected human F^V_H, object F^V_O, and the Region Of Interest (ROI) F_{ROI}, which is formed by taking the union of both candidate human and object bounding boxes. To extract these visual features, a pretrained convolutional neural network (CNN) is employed, followed by two fully connected layers. Figure 4.2 provides a visual representation of this process.

After recognizing individual objects, our visual system proceeds to analyze the relationships and interactions between them, which involves determining the spatial arrangements, relative positions, and orientations of objects in the scene [96]. In the proposed system, the spatial features related to the candidate interacting human-object pair are encoded using a spatial attention feature map. This approach is inspired by the work of [97] and [58]. As depicted in Figure 4.3, a two-channel binary image representation is employed to model the spatial relationship between a human and an object in an image. To create this representation, the bounding boxes of the human and object are merged into a reference frame, which is then resized to a fixed size. Subsequently, a binary image with two channels is generated. The first channel indicates the presence or absence of the human within its bounding box, while the second channel indicates the presence or absence of the object within its bounding box. These two-channel binary images are then fed into a two-layer convolutional neural network, enabling the extraction of a spatial attention feature map. This feature map captures information about the relative position and orientation of the human and object in the image. By utilizing this spatial attention feature map, the model can gain insights into the spatial relationship between the interacting human-object pair.

In order to represent the appearance features extracted from the image, the

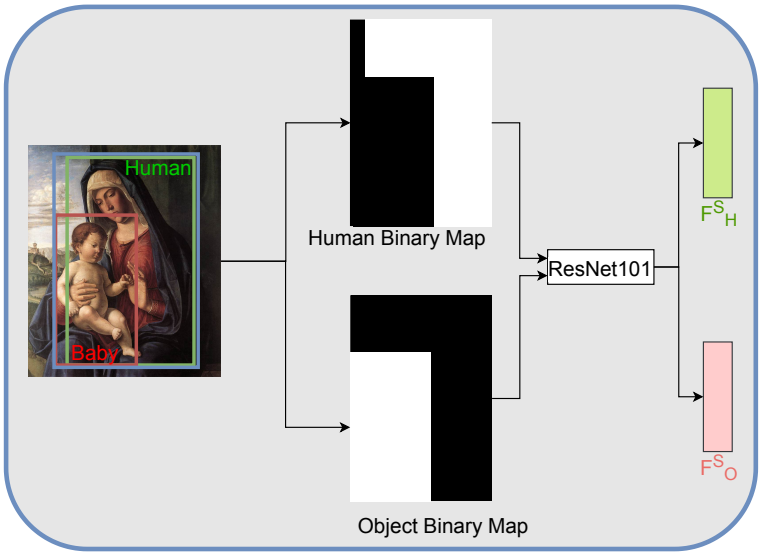


Figure 4.3: Spatial stream of the proposed system.

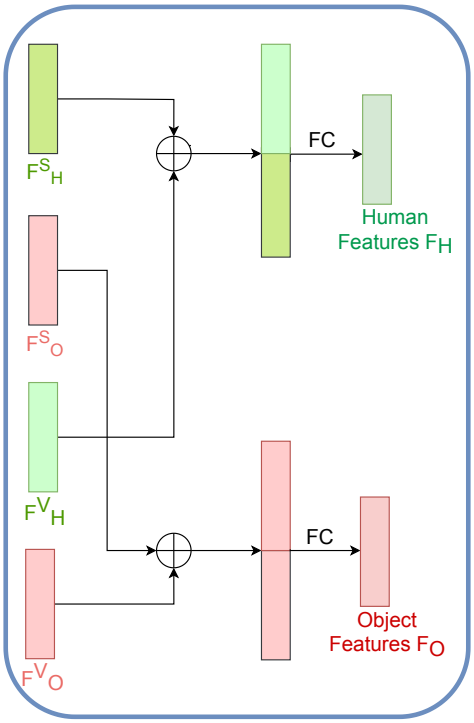


Figure 4.4: Concatenation of visual and spatial streams in the proposed system.

information from the episodic memory is integrated. This involves concatenating the human and object visual features, denoted as F_H^V and F_O^V, respectively, with the corresponding human and object spatial features, denoted as F_H^S and F_O^S. By concatenating these features, a comprehensive representation of the human and object appearances is formed. This is achieved by combining the visual information with the spatial information that captures the relative positions and orientations

of the human and object. The resulting concatenated features are then passed through a fully connected layer, which transforms them into the final human and object appearance representations, denoted as F_H and F_O. Figure 4.4 provides a visual illustration of this concatenation process, showcasing how the visual and spatial features are combined to obtain the ultimate appearance representations for both the human and object.

On the other hand, psychologically, the semantic memory in humans plays a crucial role in assessing the functional and semantic relationships between objects. This includes identifying objects as tools, understanding how objects are related in specific activities, and recognizing cause-effect relationships between objects [98]. In our perception system, contextual information and background knowledge about the detected objects are taken into account to aid in scene understanding.

To incorporate this contextual knowledge into the system, I utilize a knowledge graph based on Graph Convolution Network (GCN) in the semantic memory module. The graph consists of nodes representing all possible verbs and object categories in the annotations, while the edges connect valid pairs of verbs and object categories based on the training dataset annotations. This graph serves as a representation of the semantic relationships between verbs and object categories.

To encode the verb and object nodes in the graph, I leverage Bidirectional Encoder Representations from Transformers (BERT), which is a transformer-based model used in Natural Language Processing (NLP). BERT provides contextual word embeddings based on its training on a large corpus, allowing for a more nuanced representation of words compared to context-free models like GloVe. BERT takes in the words as input and generates output semantic features for each word. These features are then used to initialize the graph nodes for the object and verb categories, denoted as F_O^s and F_V^s respectively. The graph nodes are connected together, in an undirected graph, based on the ground truth labels from the training dataset. The graph undergoes two convolutional layers to refine and propagate information, resulting in updated semantic representations $F_O^{'s}$ and $F_V^{'s}$. These updated semantic features from the object nodes, $F_O^{'s}$, are concatenated with the visual features of the object, $F_O^{'v}$, and then combined with the updated semantic features of the candidate verbs, $F_V^{'s}$, to obtain potential interaction features $F_{Int}^{'s}$. Furthermore, at the output of the semantic stream, the semantic features of the detected object are concatenated with the semantic features of the candidate verbs to represent the interaction F_Int^s in the context of the object. This integration of semantic knowledge through the knowledge graph and BERT embeddings allows the system to capture the semantic similarity between the detected human-object pair and the candidate verbs, enhancing the understanding of the interaction in the given context.

To predict the action for each detected human, the appearance features F_H are fed into a fully connected layer followed by a Sigmoid activation function, resulting in the action prediction score s_H specifically for the human component. Similarly, the appearance features F_O and F_{ROI} representing the object and ROI, respectively, are passed through separate fully connected layers followed by Sigmoid activation functions. This yields the action prediction scores s_O and s_{ROI} for the object and ROI components, respectively.

To train the model, individual cross-entropy losses [99] are calculated for each component of the episodic stream output. The cross-entropy loss $\mathcal{L}cross^{H}$ measures the discrepancy between the predicted score s_H and the ground truth label for the human component. Similarly, $\mathcal{L}cross^{O}$ quantifies the difference between the predicted score s_O and the ground truth label for the object component, and $\mathcal{L}cross^{ROI}$ captures the difference between the predicted score $sROI$ and the ground truth label for the ROI component. These losses indicate how well the predicted scores align with the desired output for each individual component of the episodic memory module in HOI prediction.

The individual cross-entropy loss, $\mathcal{L}_{cross}$, is defined as:

$$\mathcal{L}_{cross} = -(y * log(\hat{y}) + (1 - y) * log(1 - \hat{y})) \tag{4.1}$$

where, y is the prediction score output of of the ground truth action from the sigmoid function and $\hat{y}$ is the prediction score output of all the candidate actions from the sigmoid function. By optimizing these individual cross-entropy losses, the model learns to minimize the difference between the predicted scores and the ground truth labels for each component, thus improving the accuracy of the overall HOI prediction.

Inspired by [100], the feature representations F^s_{Int} and F_{ROI} are projected into a joint embedding space using transformation matrices W_v and W_g, along with biases b_v and b_g. Specifically, the visual region features F_{ROI} are transformed into the joint embedding space as $\phi_v = W_v F_{ROI} + b_v$, while the semantic interaction features F^s_{Int} are transformed as $\phi_g = W_g F^s_{Int} + b_g$.

The objective is to maximize the cosine similarity between the learned visual and semantic representations of matching pairs, while minimizing it for non-matching pairs. This objective is achieved by training the model to minimize the graph loss $Loss_{Graph}$, which captures the discrepancy between the cosine similarities of matching and non-matching pairs:

$$\mathcal{L}_{Graph} = \begin{cases} 1 - cos(\phi_v, \phi_g) & \text{if } y = 1 \\ \max(0, cos(\phi_v, \phi_g) - \alpha) & \text{if } y = 0 \end{cases} \tag{4.2}$$

where, α is the margin and y is set to 1 if the candidate verb is the ground truth and zero if not.

During training, the model learns to adjust the embedding weights W_v and W_g, as well as the biases b_v and b_g, in order to maximize the similarity between matching pairs and minimize the similarity between non-matching pairs in the joint embedding space. By optimizing the graph loss, the model effectively learns to associate the visual and semantic features of interacting pairs, enhancing the overall performance of the HOI detection system.

The losses from the episodic memory module ($\mathcal{L}^H_{cross}$, $\mathcal{L}^O_{cross}$, $\mathcal{L}^{ROI}_{cross}$) and the semantic memory module ($\mathcal{L}_{Graph}$) are combined using a weighted sum to obtain the final loss function. The weighted sum assigns different importance to each loss component, allowing for fine-tuning of the model's behavior during training. The

final loss function for the HOI-Paint model is defined as:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}^{H}_{cross} + \lambda_2 \mathcal{L}^{O}_{cross} + \lambda_3 \mathcal{L}^{ROI}_{cross} + \lambda_4 \mathcal{L}_{Graph} \tag{4.3}$$

where $\lambda_1, \lambda_2, \lambda_3$, and λ_4 are the weight coefficients for the respective loss components. These coefficients control the relative importance of each loss term in the overall training process. By adjusting the weights, the model can be optimized to prioritize certain aspects of the HOI detection task, such as improving the accuracy of human predictions or enhancing the semantic consistency of the interactions.

During training, the model aims to minimize the total loss $\mathcal{L}_{total}$ by updating the network parameters through back-propagation. By optimizing the combined loss function, the model learns to effectively leverage both episodic and semantic memory information, leading to improved performance in human object interaction detection.

4.1.2 *Single-task learning on natural images*

The objective of this thesis is to develop a system for detecting human-object interactions in paintings. Since annotated painting datasets were not readily available, I initially focused on working with a dataset of natural images until I completed the labeling process for the large dataset of paintings.

To establish a baseline model, I adopt the system proposed by Xu et al. [33]. Their model consists of two stages: a visual-spatial stream and a semantic stream. In the visual-spatial stream, the first step involves detecting objects in the input image using a pretrained object detector, specifically Faster-RCNN [101]. Once the person and object are detected, their visual features (F^v_H and F^v_O) are extracted using a pretrained ResNet50 backbone, respectively. These visual features are then concatenated to obtain the visual representation of the region containing the human-object pair. Additionally, pairwise visual features are generated by concatenating the resulting representation with handcrafted spatial features (F_{sp}). The spatial features include the normalized distance between the bounding boxes of the human and object, as well as the logarithm of the ratio of their widths and heights. Subsequently, the final feature map is passed through two fully connected layers, and the cross-entropy loss is calculated for minimization during training.

In the semantic stream, Xu et al. utilize a knowledge graph based on Graph Convolution Network (GCN) to leverage semantic similarity and identify the verb that best describes the detected human-object pair. The graph model consists of nodes (N) representing all possible verbs and object categories in the annotations. The edges (E) in the graph connect valid pairs of verb and object category based on the training dataset annotations and an external dataset called General Visual Relationships [66]. To represent the verb and object nodes in the graph, GloVe [70] word embeddings are employed. Each object and verb word is passed through the GloVe model, generating their respective semantic feature vectors (F^s_O, F^s_V), which are used to initialize the graph nodes. The graph undergoes two convolutional layers to obtain the final semantic representations, $F^{'s}_O$ and $F^{'s}_V$.

To guide the learning of verb embeddings and exploit the semantic regularities associated with the visual modality and knowledge, joint embedding is applied. The objective is to maximize the cosine similarity between the learned transformations of the visual region and the semantic verb feature pairs, while minimizing it between non-matching pairs.

Three interventions are proposed to modify the base model, based on evidence from human psychology regarding action perception. Nelissen et al. [102] suggest that action information alone is insufficient for a complete understanding of an observed action, without knowledge about the identity of the object involved. Gallese et al. [103] state that the analysis of human movement relies on the presence of objects, as cortical responses to goal-directed actions differ when the object is present compared to when it is absent. Furthermore, Bub [104] demonstrate that observers develop specific forms of gestural knowledge derived from conceptual representations of objects, highlighting the importance of object priming in action representation.

In these networks, I apply the idea to HOI detection by priming context into the encoding of actions (*i.e.* verbs) at different levels of a deep network. I rely on the detected object's visual-spatial features as well as its semantic relationship to actions. To benefit from the influence of the object on the interaction prediction, I change the semantic representation of the actions based on their presence with the object.

My networks consists of two streams. In the first episodic memory or visual-semantic stream, features corresponding to the visual appearance, spatial features and the physical layout of people and objects are extracted as well as that of the action. The second stream is the semantic memory in which a GCN network is built between the objects and the actions. The objects and the actions are represented in the affordance-based graph by their personalized contextual vector representation extracted from a contextual word embedding model. The verb-object dependence is implemented by representing the action features as their word embeddings when they are associated with the detected object. The features from the episodic memory stream and the semantic memory stream will be used together to predict the human object interaction.

Neural Network 1 (NN1): In the first network (Figure 4.5), the external database used in the previous network for the GCN is replaced by an affordance-rich one called ConceptNet, which represents the functionality of objects. ConceptNet [105] includes data from crowd sourced resources, expert-created resources, and games such as Wiktionary which is a free multilingual dictionary and OpenCyc. To ensure that the affordance of the object is well represented in the graph, all data with the *usedfor* relationship between objects and actions are extracted, generating all possible triplets $<object, usedfor, action>$. This allows the edges in the graph to connect objects with the actions that might occur with them, based on their functionality. For example, if the detected object is *motorcycle*, the actions that are connected to it in the graph include *sit on, ride, hold, wash, clean* and the actions that have no connection to it are *eat, cook, read*. Adding these affordance based nodes enriches the graph network with nodes that help in getting better action

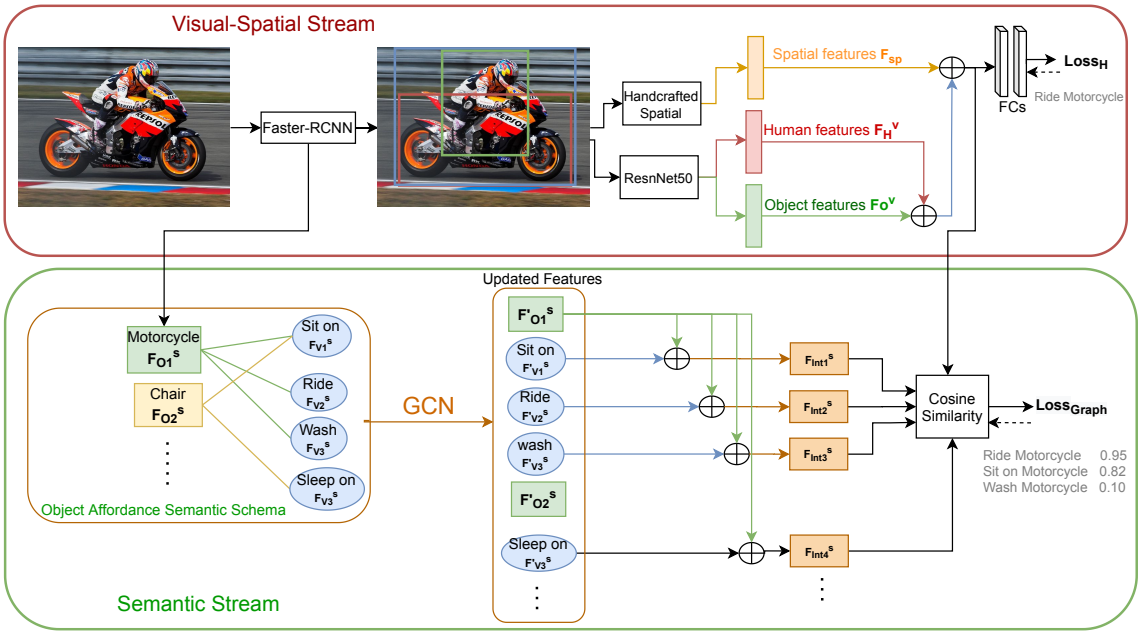


Figure 4.5: Flowchart of the first proposed HOI system NN1.

predictions.

To initialize the GCN, objects and verbs are represented using BERT [74] providing contextual word embeddings based on the large corpus it is trained on, which allows for a more nuanced representation of the words compared to GloVe. The words are fed to BERT and the output semantic features of each word is used to initialize the graph nodes. In addition, the detected object's semantic features are concatenated with that of the candidate verbs at the output of the semantic stream to represent the interaction, F_Int^s in the context of the object. The resulting interactions are then compared, using the cosine similarity, to the visual features for optimization.

Neural Network 2 (NN2): In addition to NN1, instead of concatenating the human and object features to represent the interaction, visual features, F_Int^v, are extracted from the region containing the union box of the detected human and object. The view of psychology on this matter is summarized in one simple sentence: 'The whole is more than the sum of its parts' [106]. Moreover, Baldassano et al. [86] studied the mechanism of how the brain builds HOI representation and concluded that the encoding of HOI is not simply the sum of human and object. Therefore, contextual information is incorporated through features from the union region of a human and object bounding boxes.

Moreover, estimating the pose of detected individuals, in addition to the spatial and visual features, can help improve interaction prediction by providing information about the posture of the person performing the action. To extract the human 2D body pose, denoted as F_H^p, I utilize a pretrained pose estimation model called RMPE [61] to obtain the joint positions of all detected individuals. The flowchart of NN2

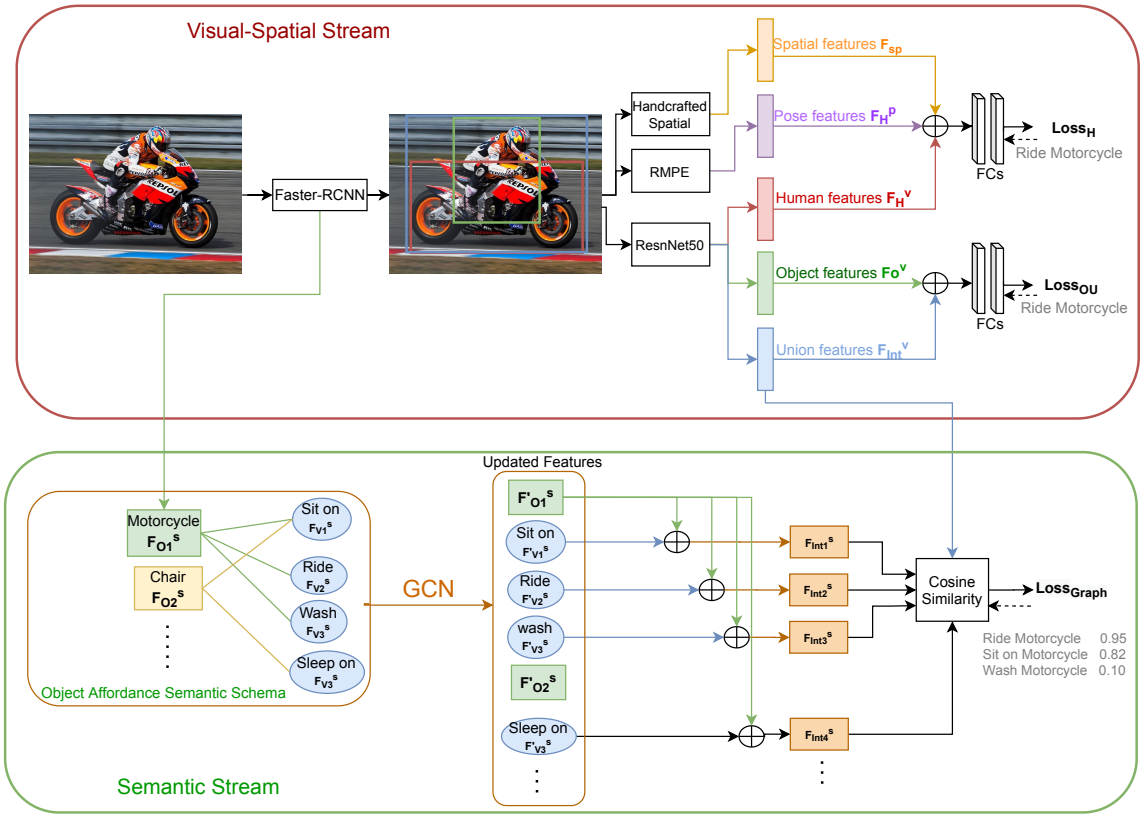


Figure 4.6: Flowchart of the second proposed HOI system NN2.

is depicted in Figure 4.6.

Neural Network 3 (NN3): In the final network, as depicted in Figure 4.7, the Faster-RCNN model is fine-tuned on the HICO-DET dataset specifically for improved object detections during testing. Furthermore, the handcrafted spatial features are replaced with a spatial attention feature map, which is created from the bounding boxes of both the human and object, following the approaches proposed in [97] and [58]. To model the spatial relationship between a human and an object, a two-channel binary image representation is utilized. The union of the two bounding boxes is taken as a reference and rescaled to a fixed size. Subsequently, a binary image with two channels is generated.

In the semantic stream, the interaction phrase composed of the verb followed by the object ($< verb, object >$) is inputted into BERT. The resulting vector representation of the verb (the first word) is extracted, considering the context of the detected object. These new representations are more specific to the detected object, tailoring the features to its characteristics. A context-based GCN is constructed using these representations. The object nodes in the GCN are represented by the semantic features of the object, while the verb nodes are represented by their semantic word embeddings. It's worth noting that in NN3, the semantic stream differs from NN2 in that object-related verbs are represented by their contextual semantic features, while non-related verbs are represented by their general semantic features.

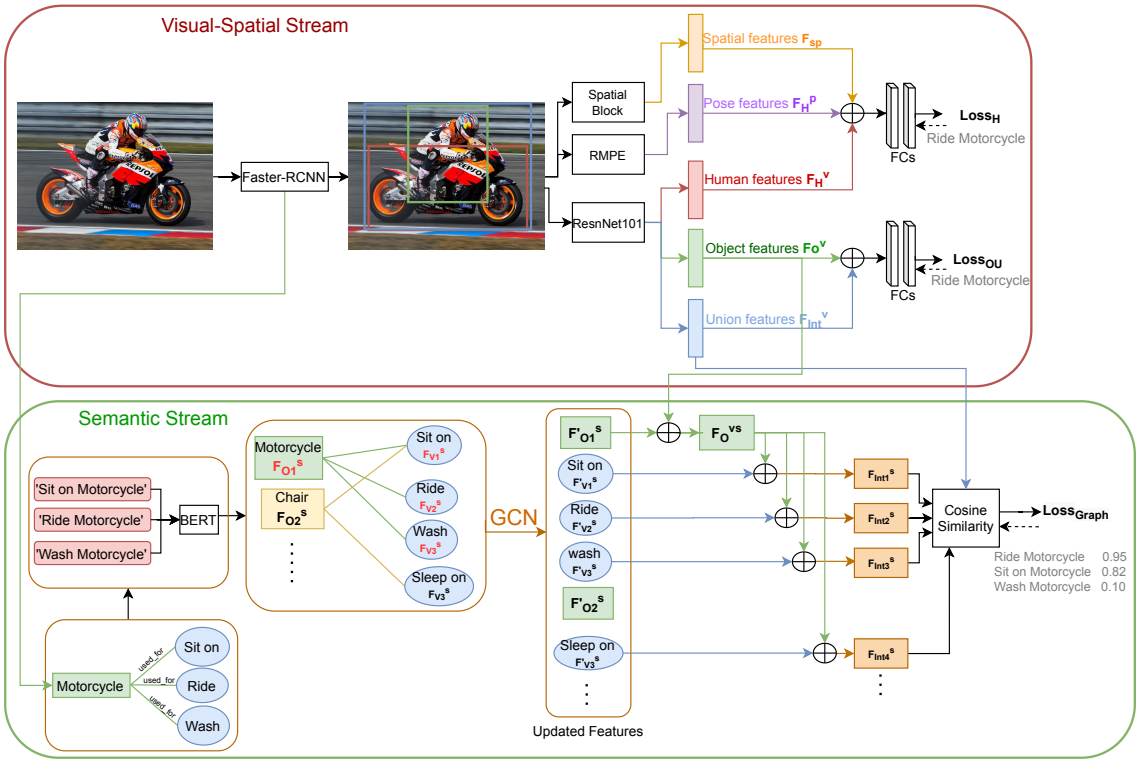


Figure 4.7: Flowchart of the third proposed HOI system NN3.

At the output of the GCN, the updated semantic features of the object are concatenated with its visual features, resulting in the final representation of the object, denoted as $F_O^{'vs}$. Subsequently, the final semantic interaction features, denoted as $F_{Int}^{'s}$, are compared to the visual features of the union box using cosine similarity to calculate the graph loss, which is used for optimization.

4.1.3 *Multi-task learning on paintings*

Multi-task learning is a machine learning approach that involves training a model to perform multiple related tasks simultaneously, rather than training separate models for each individual task. The idea behind multi-task learning is inspired by human learning and cognitive psychology [107]. In cognitive psychology [108], humans are known to learn multiple tasks simultaneously and leverage the knowledge and skills acquired from one task to help improve performance on another related task. Multi-task learning aims to capture this transfer of knowledge and leverage the shared information across tasks to improve overall performance.

For instance, many real-world problems involve multiple tasks that are inherently interrelated or share common underlying structures. By learning multiple tasks jointly [109], the model can better capture the dependencies and relationships between the tasks, leading to improved performance. Moreover, multi-task learning encourages the model to learn shared representations that are relevant to multiple tasks. This shared knowledge helps in generalizing and transferring information

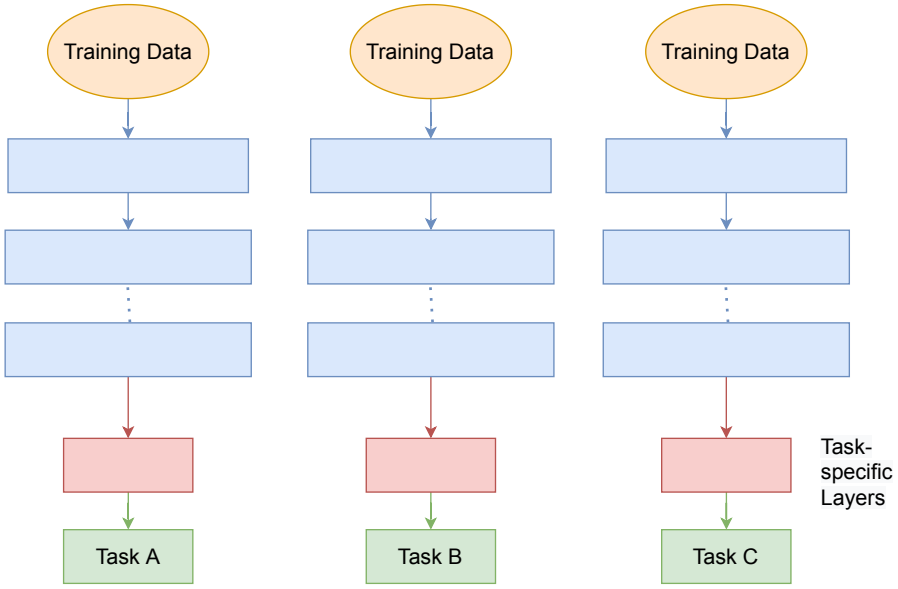


Figure 4.8: Architecture of a Single task Learning (STL) model.

across tasks, leading to improved performance on each individual task. In addition, learning multiple tasks simultaneously acts as a form of regularization, preventing overfitting on individual tasks by leveraging the shared information. This regularization effect leads to better generalization and improved performance on unseen data. Furthermore, training a single multi-task model can be more data-efficient compared to training separate models for each task. By jointly learning from multiple tasks, the model can leverage the available data more effectively, especially when the data for indivicdual tasks is limited.

Overall, the psychology behind multi-task learning aligns with the idea that learning multiple related tasks jointly can improve learning efficiency, knowledge transfer, and overall performance [110], [111]. By mimicking the way humans learn and leverage shared knowledge, multi-task learning algorithms strive to achieve similar benefits in machine learning tasks.

In artwork classification, Multi-Task Learning (MTL) models are used to jointly compute several artistic-related tasks in unison (*e.g.*, author classification, type classification, etc.) via hard parameter sharing, and obtain an aggregated loss from the losses of each independent task. By optimizing a single aggregated loss, the model is enforced to find common elements and capture relationships between the different attributes. Contextual information is provided by the painting images themselves by considering the relationships between common elements in the visual appearance of the images when multiple artistic tasks are trained together.

MTL is a subfield of machine learning in which multiple tasks are simultaneously learned by a shared model. MTL models [112] aim to solve multiple tasks jointly with the hope of generating generic features that are more powerful than those obtained through Single-Task Learning (STL) representations, where each task is trained in a separate model, as shown in Figure 4.8. Such approaches offer advantages like improved data efficiency, reduced overfitting through shared representations, and

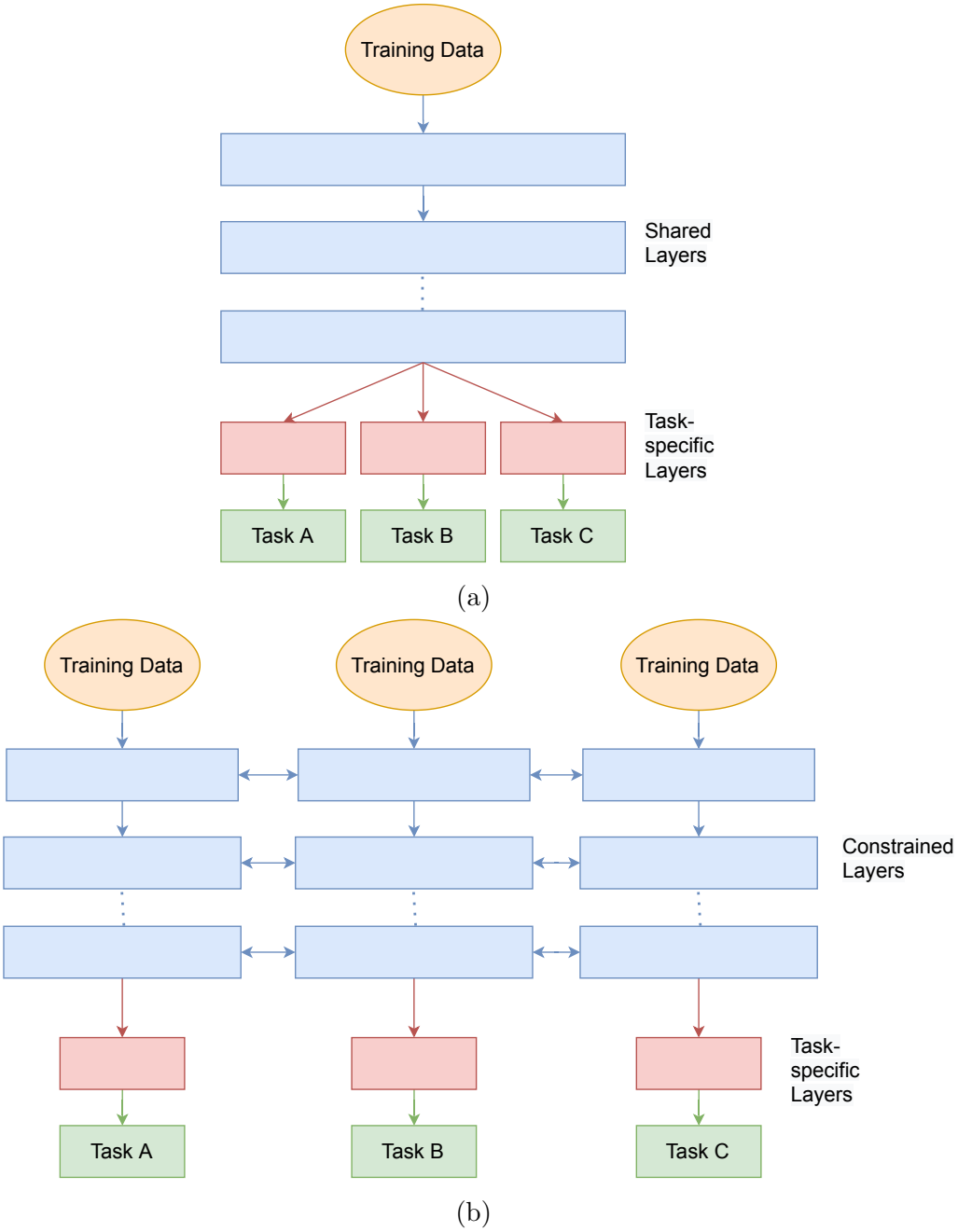


Figure 4.9: Different architectures of Multi-Task Learning models: (a) Multi-task learning with hard parameter sharing, (b) Multi-task learning with soft parameter sharing.

fast learning by leveraging auxiliary information. In deep learning approaches, MTL is commonly performed via hard or soft parameter sharing [113]. In hard parameter sharing (Figure 4.9a), parameters are shared between all the tasks, while keeping task-specific output layers for each task. In soft parameter sharing (Figure 4.9b), each task has its own model with its own parameters and the distance between the parameters of the model is then regularized in order to encourage the parameters

to be similar.

OmniArt, proposed by Strezoski et al. [114], is a popular multi-task method used for artwork classification. This approach uses a multi-output CNN model that employs a shared convolutional base, ResNet50, for feature extraction, and separate output layers for each task. The model is trained by minimizing an aggregated loss, which is obtained by combining the separate losses using a weighted combination. The use of a shared base ensures that the model captures common features across different tasks, which leads to better generalization and efficiency.The model is evaluated on the Rijks'14 dataset, which consists of 112,039 images classified according to multiple attributes, including 6,626 artists, 1,054 types, 406 materials, and 628 periods during which they were created. Additionally, the authors introduce their own dataset, OmiArt, containing 432,217 artworks classified based on multiple attributes, including 21,364 artists, 837 types, created on 6,385 materials, and during 2,389 periods. The results show that the multi-task approach outperforms single-task models and achieves good performance across all the different classification tasks. Furthermore, the authors also demonstrate the importance of using a shared base by comparing the performance of their multi-task model with and without a shared base.

Belhi et al. [115] proposed a multi-modal architecture that combines both digital images and textual metadata to classify the artist and the creation year in paintings. The model takes a three-channel image input and passes it through the convolutional base of a standard ResNet50. The textual metadata, including genre, medium, and style, are one-hot-encoded and provided as input to a shallow feedforward network. The visual and textual features are then concatenated and used to feed the final classification layer. The authors collected paintings from three different datasets: the WikiArt dataset (10,000 paintings), the METropolitan Museum of New York (MET) dataset (20,000 paintings), and the Rijksmuseum dataset (13,594 paintings), totaling 43,594 paintings annotated for the artist name, year of creation, genre, style, and medium. The experimental results showed that the multi-modal classification system outperformed the individual classification in most cases, demonstrating the effectiveness of combining multiple sources of information for artwork classification.

Bianco et al. [7] propose a new approach to solving the tasks of artist, style, and genre categorization in a multi-task model. Their proposed model is a deep multi-branch neural network that takes crops of the input image at different resolutions to gather information from low-level texture details and exploit the coarse layout of the painting. The authors compare two different cropping strategies: a random strategy and one based on Spatial Transformer Networks (STN). Additionally, they experiment with injecting different hand-crafted features directly computed on the input images, such as Local Binary Patterns (LBP), Generalized Search Trees (GIST), histogram of oriented gradient (HOG), and others. The evaluation is carried out on a new dataset built by the authors, named MultitaskPainting100k, sourced from wikiart.org. The dataset consists of 100K paintings divided into 1508 artists, 125 styles, and 41 genres. The authors find that using the STN cropping strategy and injecting HOG features achieves the best accuracies for style, genre, and artist classification. They also demonstrate that applying multi-task learning improves the

accuracies on the three tasks compared to applying single-task learning for each task separately.

The approach proposed by Garcia et al. [116] is a hard parameter sharing MTL method that aims to obtain context-aware embeddings for art analysis. By jointly learning multiple artistic tasks, the resulting visual representations are enforced to capture relationships and common elements between four different artistic attributes: author, school, type, or period, providing contextual information about each painting. They use a standard common CNN, ResNet50, to extract deep features from each input image and feed them to separate fully connected layers for classification. They test their model on a multi-modal dataset for semantic art understanding called SemArt, which contains 21,384 fine-art images, each with its respective attributes, as well as a short artistic comment or description. The paintings are classified according to 10 different common types, 25 schools, 18 different timeframes, and 350 authors. Through their experiments, they showed that MTL was able to enhance the art classification accuracies for all tasks compared to training a separate model for each task.

Zhao et al. [117] propose a novel approach to classify paintings' type, school, timeframe, and author using a graph convolutional network (GCN) and artistic comments from the SemArt dataset. Instead of relying solely on painting visual features, they incorporate natural language processing (NLP) techniques to extract information from the comments. They construct a single artistic comment graph based on co-occurrence relations and document word relations, and train an ArtGCN on the entire corpus. The model uses four different GCN layers in the second layer to address the various tasks. Their results demonstrate that this approach outperforms previous state-of-the-art methods for the task of joint art classification.

ArtSAGENet is a multimodal architecture proposed by Efthymiou et al. [118], which integrates Graph Neural Networks (GNNs) and CNNs to jointly learn visual and semantic-based artistic representations. The authors represent the nodes in the graph using GraphSAGE [119], a general inductive framework that leverages node feature information (such as text attributes) to efficiently generate node embeddings for previously unseen data. The model is tested on a large dataset of 75,921 paintings from the WikiArt dataset and is used to classify them according to 750 artists, 20 styles, and 13 timeframes.

Yang et al. [120] proposed an adaptive multi-task learning method for automatic art classification tasks. They used a Lagrange multiplier strategy to weight multiple loss functions within the multi-task learning framework, enabling the system to adaptively learn the weights of each task. The proposed method was tested on the SemArt dataset [121], where the authors jointly solved four tasks: type, school, timeframe, and author. Their experiments demonstrated that the adaptive multi-task learning method improved the performance of art classification compared to traditional multi-task learning methods.

Castellano et al. [122] presented ArtGraph, an artistic Knowledge Graph (KG) based on WikiArt and DBpedia. These nodes were interconnected, and each of them was connected to its attributes. For example, the artist's attributes included the field they belonged to, the movement they belonged to, and others. The artwork's

Table 4.1: Performance of the aforementioned models in STL and MTL.

Model	Setting	Dataset	Task Accuracy (%)						
			Artist	Style	Genre	School	Type	Time	Material
[114]	MTL	Rijks'14	**52.2**				**93.7**	70.1	**98.0**
	STL		50.3				91.7	**71.2**	97.2
	MTL	OmniArt	**64.5**				**99.4**	77.9	**76.8**
	STL		60.7				99.0	**79.3**	74.0
[7]	MTL	Multitask Painting100k	56.5	57.2	63.6				
[116]	MTL	SemArt	**60.3**			**69.1**	**79.1**	**61.6**	
	STL		55.7			63.6	78.7	59.2	
[117]	MTL		**61.5**			**66.7**	**79.0**	**61.6**	
	STL		52.6			63.5	77.1	59.2	
[120]	MTL		**61.5**			**78.7**	**80.5**	**65.5**	
[118]	MTL	WikiArt	**76.6**	**77.6**				**79.2**	
	STL		65.5	70.1				71.4	
[122]	MTL	WikiArt + DBpedia		**58.58**	**76.13**				
	STL			58.31	71.23				

attributes included the media it was made of, its style, genre, and others. This graph provided knowledge discovery capabilities without the need to train a learning system. The authors used ArtGraph to build a KG-enabled painting classification method. Their method extracted embeddings from ArtGraph and injected them as contextual knowledge into a deep learning model. A concatenation layer received both the contribution of visual embeddings, extracted from a Vision Transformer (ViT), and graph embeddings extracted from ArtGraph using a Graph Attention Network (GAT), respectively. The overall network learned to minimize the error made in predicting the correct style and genre of a given input painting.

One of the key advantages of MTL models is that they allow for the optimization of a single aggregated loss that combines the losses of each individual task. By doing so, the model is able to focus on finding common elements between the tasks and identifying relationships between them. This is particularly useful in the context of artwork classification, where different artistic attributes may be interrelated and correlated, such as the relationship between the style and the artist.

Another advantage of MTL models is that they are able to incorporate contextual information provided by the painting images themselves. This is achieved by considering the relationships between common elements in the visual appearance of the images when multiple artistic tasks are trained together. For example, if the model is simultaneously trained to classify the artist and the style of a painting, it may learn to identify common visual elements that are characteristic of the artist's style, such as brushstrokes, color palette, and composition.

MTL models have proven to be effective in artwork classification, allowing for the joint computation of multiple artistic-related tasks and capturing relationships be-

tween different attributes, while also taking into account the contextual information provided by the images themselves.

Table 4.1 presented the prediction performance of the aforementioned models under MTL and STL settings. Most of the results showed that tasks, when learned together in a single model, tended to perform better compared to when trained separately. However, the test results presented by [114] and [117] indicated that combining all the tasks together in a single model was not always the best solution to improve the model's performance. For instance, in [114], the Timeframe classification was more accurate when the model was trained in a STL setting. This meant that some tasks could have a negative effect when trained with others due to conflicting needs. In this case, increasing the performance of a model on one task would hurt performance on a task with different needs, a phenomenon referred to as negative transfer or destructive interference. Minimizing negative transfer was a key goal for MTL methods.

Hence, the challenge is to identify the optimal combination of tasks to improve HOI, but testing all possible permutations is computationally expensive. Recently, a paper from Stanford [123] proposed various approaches to reduce the search for the best combination of tasks. One of the most effective proposed methods is the Higher-Order Approximation (HOA) from lower-order, which suggests that a simple average of the first-order networks' training losses is a good estimator of the training loss of a higher-order network. By using this strategy, one can predict the performance of all networks with three or more tasks using the performance of all fully trained two-task networks. Firstly, all networks with two or fewer tasks are trained to convergence. Then the performance of higher-order networks is predicted, network selection is performed on both the trained and predicted networks, and the higher-order networks are trained from scratch.

To further investigate the HOA theory, I took the model of [116] and applied HOA to solve the painting classification based on author, type, school and timeframe. The images are fed to ResNet50 without its last fully connected layer to extract the shared visual embeddings. Then, the generated embeddings are input to four separate classifiers for task classification where each classifier is composed of a fully connected layer followed by a ReLU nonlinearity. The outputs from each task-specific layer are jointly used to compute a classification loss. ResNet50 is initialised with its standard pre-trained weights for image classification, whereas the weights from the rest of the layers are initialised randomly. The images are scaled down to 256 pixels on each side and randomly cropped into 224 $\times$ 224 patches. During training, the visual data is augmented by randomly flipping the images horizontally. The size of the embeddings produced by ResNet50 is 1x2048. Stochastic gradient descent with a momentum of 0.9 and a learning rate of 0.001 is used as the optimizer. The training is conducted using mini-batches of 28 samples for a maximum of 300 epochs, with a patience of 30 epochs; this means that if the calculated validation loss does not decrease for 30 epochs, the training stops. The loss weight for all tasks is set to 0.25.

I first trained every pair of tasks using the mentioned setup and report the training loss for each task when trained in each model in Table 4.2 . With four

Table 4.2: The training loss of the four tasked trained in all possible pairs.

			Training Loss of			
			Type	School	Time	Author
Trained in	2TL	TY+SC	0.0058	0.0087		
		TY+TI	0.0066		0.0086	
		TY+AU	0.0118			0.0799
		SC+TI		0.0093	0.0098	
		SC+AU		0.016		0.0695
		TI+AU			0.0133	0.0657

Table 4.3: The estimated training loss of each task when trained in the five different high order settings (3TL,4TL) based on HOA.

			Based on HOA, Training Loss of			
			Type	School	Time	Author
Trained in	3TL	TY+SC+TI	**0.0062**	**0.009**	**0.0092**	
		TY+SC+AU	0.0088	0.01235		0.0747
		TY+TI+AU	0.0092		0.01095	0.0723
		SC+TI+AU		0.01265	0.01155	**0.0676**
	4TL	TY+SC+TI+AU	0.00807	0.0113	0.01057	0.0717

different tasks, six models are trained C_2^4, each with two tasks (2TL): Type & School (TY+SC), Type & Timeframe (TY+TI), Type & Author (TY+AU), School & Timeframe (SC+TI), School & Author (SC+AU), and Timeframe & Author (TI+AU).

After training all the possible pair of tasks, I apply HOA to estimate the loss of each task in a higher order model; 3-Task Learning (3TL) and 4-Task Learning (4TL). For example, The loss of Type when trained with School and Timeframe is: (0.0058+0.0066)/2. The loss of type when training with School, Timeframe, and Author is: (0.0058+0.0066+0.0118)/3. There are, C_3^4, four different combinations of 3TL models and one 4TL model. Table 4.3 shows the estimated losses for each task in all five high order model settings. We can see that Type, School and Timeframe report the lowest training loss when trained together. While Author has the lowest training loss when trained with School and Timeframe. Therefore, to get the best model with the lowest Type, School, and Timeframe losses, one needs to train and test these three task jointly in a single model. Whereas for the Author task, one trains the MTL model consisting of School, Timeframe, and Author, but during test time, only the results for the Author are inferred. This way, The School and the Timeframe are helping the Author by incorporating their features in the shared module.

To confirm this theory, I trained all the four 3TL and the 4TL models and report the training losses for each task in five different MTL settings. The results from Table 4.4 confirm Standley et al.'s [123] higher order approximation theory, which suggests finding the optimal task combination for achieving the lowest loss

Table 4.4: The reported training loss of each task when trained in the five different high order settings (3TL,4TL).

			Training Loss of			
			Type	School	Time	Author
Trained in	3TL	TY+SC+TI	**0.0125**	**0.028**	**0.0199**	
		TY+SC+AU	0.0224	0.0458		0.2371
		TY+TI+AU	0.02		0.0232	0.2242
		SC+TI+AU		0.0419	0.0233	**0.1881**
	4TL	TY+SC+TI+AU	0.0334	0.0746	0.0385	0.3739

			Test Accuracy of			
			Type	School	Time	Author
Trained in	STL	Type	0.791			
		School		0.651		
		Time			0.598	
		Author				0.558
	2TL	TY+SC	0.775	0.649		
		TY+TI	0.785		0.609	
		TY+AU	0.779			0.537
		SC+TI		0.662	0.595	
		SC+AU		0.652		0.535
		TI+AU			0.616	**0.562**
	3TL	TY+SC+TI	<u>0.780</u>	<u>0.667</u>	<u>0.594</u>	
		TY+SC+AU	0.782	0.670		0.536
		TY+TI+AU	0.789		**0.627**	0.537
		SC+TI+AU		0.662	0.613	<u>0.543</u>
	4TL	TY+SC+TI+AU	**0.803**	**0.681**	0.606	0.528

Table 4.5: The test accuracies of all four tasks under the different STL, 2TL, 3TL, and 4TL settings.

and optimal model performance.

As mentioned by [123], HOA comes with a penalty in terms of prediction quality. However, this technique requires only a quadratic number of networks to be trained rather than an exponential number, and would theoretically be advantageous when the number of tasks is large. In Table 4.5, the test accuracies are reported for the four tasks trained in all the possible combinations. The highest accuracy achieved for each task is indicated in bold. Meanwhile, the underlined number represents the test accuracy for the four tasks in the setting determined by the HOA.

Another challenge in MTL is appropriately weighting the loss functions for each task. A common approach to facilitate multi-task optimization is to balance the individual loss functions across different tasks. When training a model on multiple tasks, the task-specific loss functions need to be combined into a single function that the model is trained to minimize. This raises the question of how to design

a combined loss function that is suitable for all tasks. Hence, it is necessary to consider multiple loss weighting strategies during the design of the MTL model to achieve optimal performance across all tasks. These strategies, such as Gradient Normalization (GradNorm), Uncertainty Weights (UW), Conflict-Averse Gradient Descent (CAGrad), Random Loss Weight (RLW), and others (see Table 4.6), aim to balance the different tasks and mitigate negative effects.

Table 4.6: Loss weighting strategies used in multi-task learning.

Weighting Strategy
Equal Weighting (EW)
Gradient Normalization (GradNorm) [124]
Uncertainty Weights (UW) [125]
Homoscedastic Uncertainty (HW) [126]
Dynamic Weight Average (DWA) [127]
Geometric Loss Strategy (GLS) [128]
Projecting Conflicting Gradient (PCGrad) [129]
Gradient sign Dropout (GradDrop) [130]
Impartial Multi-Task Learning (IMTL) [131]
Gradient Vaccine (GradVac) [132]
Conflict-Averse Gradient descent (CAGrad [133]
Random Loss Weighting (RLW) [134]

To enhance the detection of human-object interactions in paintings, I propose the application of multi-task learning to improve the model's learning capability. Within the multi-task learning framework, four tasks are learned simultaneously: artist classification, timeframe estimation, type recognition, and author classification. This simultaneous learning facilitates knowledge transfer and enhances the accuracy of HOI detection. By sharing representations between tasks, the model can leverage common patterns and features, leading to improved generalization and enhanced performance in HOI detection in paintings. Furthermore, the inclusion of contextual information from the painting images, considering the relationships between shared visual elements during the training of multiple artistic tasks, enriches the model's understanding and further improves the accuracy of HOI detection.

Specifically, I augment the HOI detection model by incorporating four additional tasks related to the painting, training the model on these tasks alongside the primary HOI task. By doing so, I integrate additional attributes about the painting into the model's prediction process. Simultaneously learning multiple attributes about the painting enables the model to develop a more comprehensive understanding of the scene, significantly enhancing its ability to accurately detect human-object interactions. This is achieved by combining the supplementary information learned from the auxiliary tasks with the visual information, thereby extracting more robust and meaningful features. Consequently, multi-task learning plays a crucial role in significantly improving the performance of the HOI detection model by providing a more comprehensive and holistic representation of the painting.

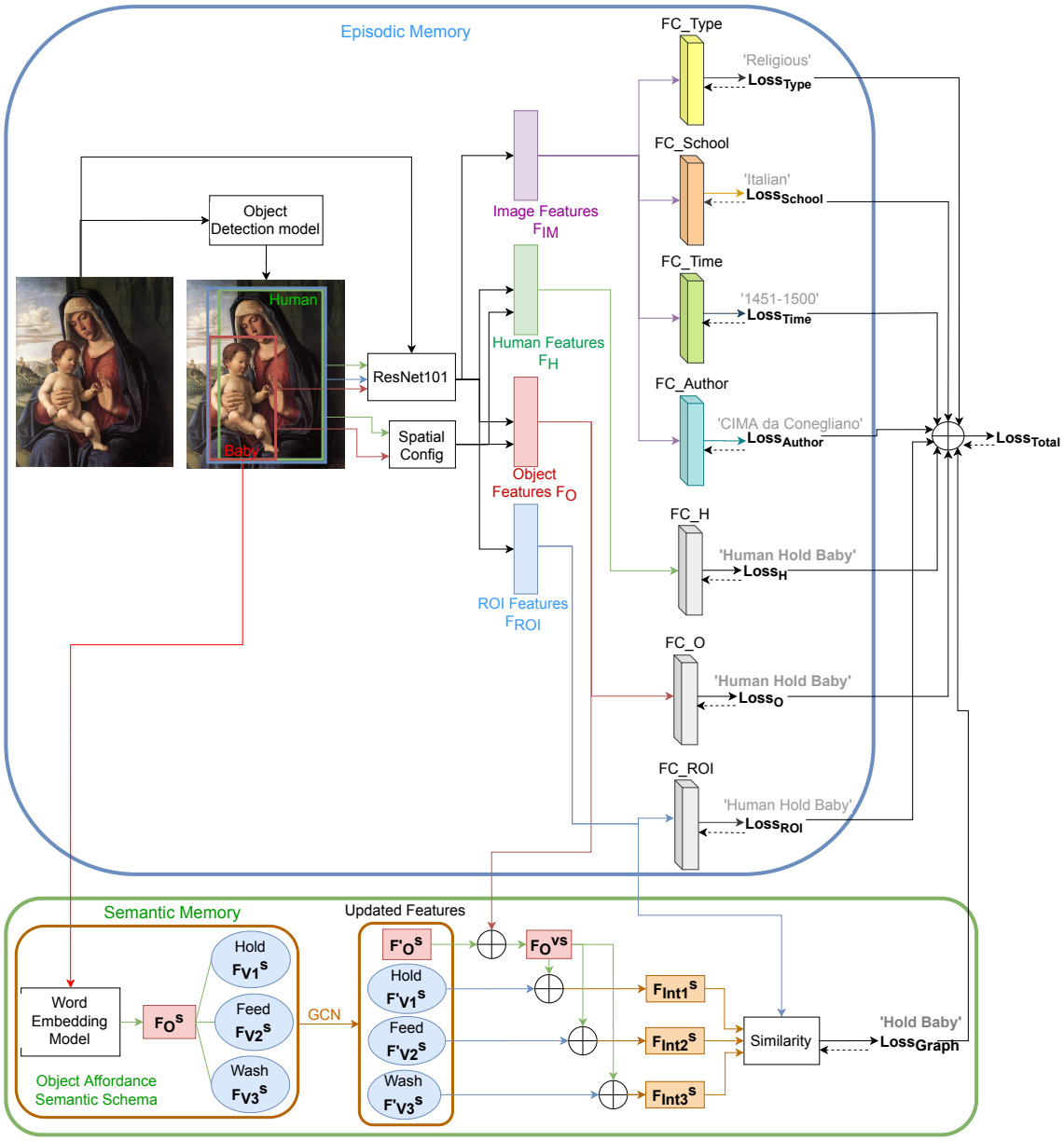


Figure 4.10: The proposed system for Human Human Object Interaction Detection using multi-task Learning **HOI-Paint-MTL**. The painting is first input to an object detection model for instance detection. In episodic memory module, appearance features from the image, the candidate human, object, and ROI are extracted using an extraction CNN (ResNet-101). In the semantic memory module, semantic features are extracted for the detected object and all the related candidate actions. The input features of the detected object-related actions are replaced with their contextual ones. Losses are calculated from the episodic and semantic modules and joint together for optimization.

The flowchart of my multi-task learning model **HOI-Paint-MTL**, illustrating the steps for the simultaneous prediction of multiple attributes in an image, is de-

picted in Figure 4.10. Following a similar approach to the single-task learning model, the initial stage involves the detection of objects and humans within the image. Utilizing a shared pretrained CNN, visual features are extracted from various regions of interest, including the cropped human ($F^V H$), object ($F^V O$), ROI (F_{ROI}), and the entire image (F_{IM}). Additionally, spatial features ($F^S H$ and $F^S O$) pertaining to the human and object are obtained and concatenated with their respective visual features, resulting in the extraction of appearance features for both the human (F_H) and the object (F_O). Subsequently, the obtained representations are passed through four separate fully connected layers, each equipped with an activation function for task classification. This comprehensive approach enables the simultaneous prediction of multiple attributes in the image, leveraging shared visual features to enhance both accuracy and efficiency. The semantic memory stream of the proposed MTL model is similar to that of the STL model, where it is only used in the interaction prediction.

In the MTL setting, the HOI classification stream closely resembles the one in the STL setting, except for the parameter sharing aspect. The distinguishing factor is that multiple tasks to be classified share parameters from the feature extraction backbone. This parameter sharing facilitates the utilization of shared representations across tasks, enabling the model to leverage visual features that are pertinent to multiple tasks concurrently. Through the exploitation of shared parameters, the MTL approach enhances performance on the HOI task resulting in a more comprehensive comprehension of the input image.

Similarly to the STL model HOI-Paint, in the episodic stream, the action prediction scores (s_H, s_O, and s_{ROI}) are computed by passing F_H, F_O, and F_{ROI} through separate fully connected layers, followed by a Sigmoid activation function. To quantify the dissimilarity between the predicted scores and the ground truth labels for each individual component of the HOI, individual cross-entropy losses are computed: $\mathcal{L}cross^H$ for the human, $\mathcal{L}cross^O$ for the object, and $\mathcal{L}_{cross}^{ROI}$ for the ROI. These losses capture the disparity between the predicted scores and the ground truth labels for each specific HOI element. Moreover, in the semantic stream, the graph loss $mathcalL_{Graph}$ defined by the cosine similarity between the semantic and visual representations of the interaction is calculated.

Moreover, the classification scores from the additional tasks, s_{Type}, s_{School}, $s_{Timeframe}$, and s_{Author}, are calculated by feeding the images features to fully connected layers followed by a sigmoid activation function. After calculating the scores the individual cross entropy losses, $\mathcal{L}_{cross}^{Type}$, $\mathcal{L}_{cross}^{School}$, $\mathcal{L}_{cross}^{Timeframe}$, and $\mathcal{L}_{cross}^{Author}$ are found.

The resulting total MTL loss, $\mathcal{L}^{Total}$, is defined as the weighted sum of the individual losses for each task:

$$\begin{aligned}\mathcal{L}^{Total} = \lambda_1 \mathcal{L}_{cross}^{H} + \lambda_2 \mathcal{L}_{cross}^{O} + \lambda_3 \mathcal{L}_{cross}^{ROI} + \lambda_4 \mathcal{L}_{Graph} \\ + \lambda_5 \mathcal{L}_{cross}^{Type} + \lambda_6 \mathcal{L}_{cross}^{School} + \lambda_7 \mathcal{L}_{cross}^{Timeframe} + \lambda_8 \mathcal{L}_{cross}^{Author}\end{aligned} \tag{4.4}$$

where, $\lambda_1, \lambda_2,, \lambda_3$, and λ_4 are added weights to each HOI task-related loss function and $\lambda_5, \lambda_6, \lambda_7$, and λ_8 are the weights to each of the additional four tasks. The final target is to minimize the total loss term in (4.4).

Chapter 5

Experiments and Results

This chapter offers an overview of the HOI detection datasets in natural images used to evaluate the models, along with the experiments conducted to validate the proposed system's performance compared to the state-of-the-art model. The chapter also presents the new HOI detection dataset, highlighting its characteristics. It also describes the experiments conducted to assess the proposed system's performance, including the metrics used for evaluation.

5.1 HICO-DET and V-COCO Datasets

Different dataset have been used to perform experiments for human object interaction detection (Table 5.1). Humans Interacting with Common Objects (HICO) dataset was first introduced by Chao et al. [26]. The HICO dataset comprises 47,774 images, encompassing 600 categories of human-object interactions denoted as $< verb, object >$. It includes 117 action classes (including one "no interaction" class) and 80 object labels. Subsequently, the HICO dataset was augmented to HICO-DET [58] by incorporating ground truth detected human and object bounding boxes. HICO-DET consists of 38,118 training and 9,658 testing images. In each image, a person can engage in multiple interactions with different objects. The training set of HICO-DET contains 117,871 interaction annotations for 600 interaction classes, while the test set includes 33,405 interaction annotations.

Table 5.1: The different benchmarks for classifying and detecting human-object interactions (HOI) in images. Nb.: Number of

Dataset	Year	Nb. Images	Train/Val	Test	Nb. Objects	Nb. Actions	Task
HICO	2015	47,776	38,118	9,658	80	117	HOI Classification
HICO-DET	2018	47,776	38,118	9,658	80	117	HOI Detection
V-COCO	2015	10,346	5,400	4,946	80	29	HOI Detection

Table 5.2: Occurrence of some of the 600 interactions of HICO-DET train set.

Interaction	Train_occurrence
ride skateboard	1456
straddle motorcycle	1030
hold dog	304
exit airplane	30
wash wine_glass	1

Table 5.3: Occurrence of some of the 117 verbs of HICO-DET train set.

Verb	Train_occurrence
hold	13998
ride	8692
kick	249
peel	53
zip	1

Tables 5.2 and 5.3 present the occurrence of a selection of verbs and interactions from the HICO-DET dataset. Interaction categories with fewer than 10 training samples are categorized as "rare," while the remaining categories are referred to as "non-rare." In total, there are 138 rare categories and 462 non-rare categories. Observing Table 5.3, it can be noted that the verb 'hold,' which appears in more than 10% of the interactions, has the highest occurrence count in the training set. Conversely, verbs such as 'zip' occur only once in the training set. This imbalance in the HICO-DET dataset is evident. Furthermore, it is worth mentioning that multiple humans may be present in the same image, but not all of them are assigned interaction labels. Additionally, some images possess interaction labels despite lacking visible human body parts.

Verbs in COCO (V-COCO) [135] is another dataset commonly used for evaluating HOI models. It is derived from the MS-COCO dataset and consists of 2,533, 2,867, and 4,946 images for training, validation, and testing, respectively. The dataset includes 16,199 instances of humans and 80 object labels. Additionally, V-COCO incorporates 29 verb labels, with 25 representing interactions with objects and 4 representing body motions such as 'run' and 'walk'. Similar to HICO-DET, the V-COCO dataset allows for multiple interactions between a person and different objects within each image.

5.1.1 *Evaluation Metrics*

Following the work of Chao et al. [58], HOI detection systems evaluate their performance on both datasets using the role mean average precision (mAP). An HOI detection is considered a True Positive (TP) if it accurately localizes the human and object (*i.e.*, the predicted box has an Intersection over Union (IoU) ratio greater than 0.5 with the ground truth) and predicts the interaction label correctly. Other-

wise, it is considered a False Positive (FP):

$$Detection = \begin{cases} TP, & \text{if } IoU(pred\_BB, GT\_BB) \geq 0.5 \\ & \&pred\_Action = True \\ FP, & \text{otherwise} \end{cases}$$

There are two modes of mAP evaluation for HICO-DET: the Default (DT) mode and the Known-Object (KO) mode. In the DT mode, each HOI category is evaluated on all testing images, while in the KO mode, an HOI is only evaluated on images that contain its associated object category. The mAP is reported for three different category sets: (1) all 600 HOI categories in HICO (Full), (2) 138 HOI categories with less than 10 training instances (Rare), and (3) 462 HOI categories with 10 or more training instances (Non-Rare).

In V-COCO, there are HOIs defined without object labels. To handle this situation, the performance is evaluated in two different scenarios based on V-COCO's official evaluation scheme. In Scenario 1, detectors are required to report cases where there is no object, while in Scenario 2, the prediction of an object bounding box is ignored in these cases.

5.2 SemArt-HOI Dataset

The task of human object interaction detection in paintings is a relatively new research area that has not received much attention in the past in no small part due to the lack of suitable datasets. To overcome this limitation, I propose the SemArt-HOI dataset, which is the first dataset specifically designed for human object interaction detection in paintings. I augment the existing SemArt dataset with object detections and interaction labels, allowing training and evaluating HOI detection models more effectively.

The SemArt dataset, introduced by Garcia and Vogiatzis [121], is a comprehensive multi-modal dataset designed for semantic art understanding (https://researchdata.aston.ac.uk/id/eprint/380/). It consists of a diverse collection of 21,384 fine-art images, each accompanied by corresponding attributes such as Type, School, Timeframe, and Author. Additionally, each painting is associated with a concise artistic comment or description. The paintings in the SemArt dataset are classified into 10 different common types, including portrait, landscape, religious, study, genre, still life, mythological, interior, historical, and more. They are also categorized into 25 distinct schools, such as Italian, Dutch, French, Flemish, German, Spanish, English, and others. The dataset further covers 18 different timeframes, representing periods of 50 years each, ranging from 801 to 1900. Moreover, SemArt includes information about 350 different authors, including renowned artists like Vincent van Gogh, Claude Monet, Giovanni Santi, Michelangelo Cerquozzi, Nicolas Poussin, and many more.

To annotate the SemArt-HOI dataset, I utilized the Make Sense tool [136], which is an online platform that offers free photo labeling capabilities. Using this tool,

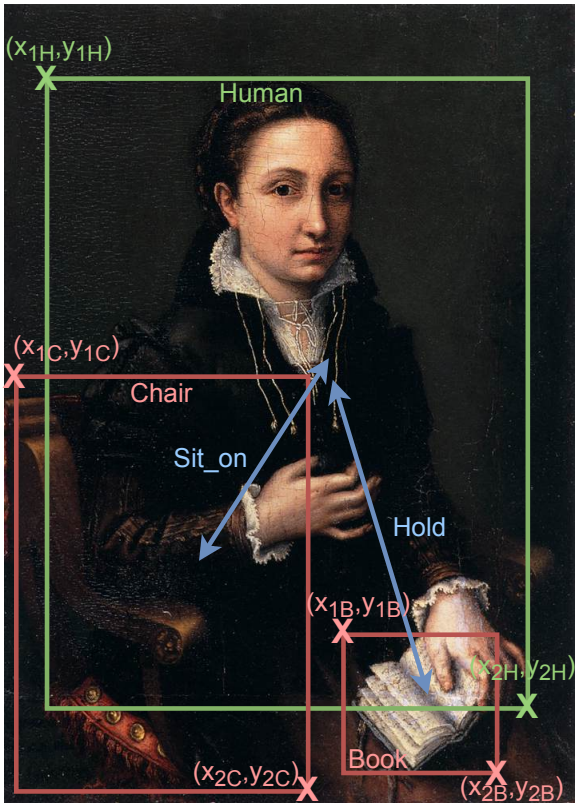


Figure 5.1: The human object interaction annotation on the SemArt-HOI dataset. The dataset includes objects such as humans, chairs, and books, which are detected and assigned bounding boxes. Each bounding box is defined by the coordinates of its top left corner (x_1, y_1) and bottom right corner (x_2, y_2). Additionally, an interaction verb label, such as "sit_on" or "hold," is assigned to each pair of a human and an object in the image.

I annotated every image in the dataset with the corresponding Human-Object-Interaction (HOI) triplet, represented as $< human, predicate, object >$. The annotation process includes drawing bounding boxes around each human and object involved in an interaction within the image. For each instance of interaction, I defined a bounding box using its upper-left corner coordinates (x_1, y_1) and bottom-right corner coordinates (x_2, y_2). Additionally, I assigned a class label to each bounding box to indicate the object category, such as human, chair, book, and so on.

In the second step of the annotation process in each image one draws an arrow to visually connect the corresponding entities, representing the interaction between them. Additionally, an interaction verb class, such as "sit_on" or "hold," is assigned to each pair, providing a specific label for the observed interaction. Figure 5.1 illustrates an example of this annotation process. By completing this second step, I finalized the annotation process for the SemArt-HOI dataset, ensuring that all necessary information for training and evaluating HOI detection models was included. The SemArt dataset consists of 14,187 images out of the original 21,384 images that contained pairs of interacting entities. The remaining images either lacked human

Table 5.4: SemArt and SemArt-HOI datasets.

	SemArt	SemArt-HOI
Number of Images	21,384	14,187
Type	10	10
School	25	25
Timeframe	18	18
Author	350	343
Object	-	99
Verb	-	66
Interaction	-	248

presence or depicted humans that were not engaged in any interactions with objects. Consequently, the SemArt-HOI dataset consists of these 14,187 labeled images.

The SemArt-HOI dataset encompasses various categories, including 10 different types, 25 schools, 18 timeframes, and 343 authors, as displayed in Table 5.4. Each image within the dataset features a minimum of two instances of 99 objects, including humans, chairs, dogs, horses, tables, and beds. Furthermore, each image contains at least one interaction verb from a selection of 66 verbs, such as hold, sit_on, write_on, and paint. In total, the SemArt-HOI dataset encompasses 248 unique interactions.

The SemArt-HOI dataset contains a significant amount of object and interaction annotations, with a total of 73,683 object bounding box annotations and 40,101 interaction labels. To provide a better understanding of the distribution of objects and interactions in the dataset, several charts have been created.

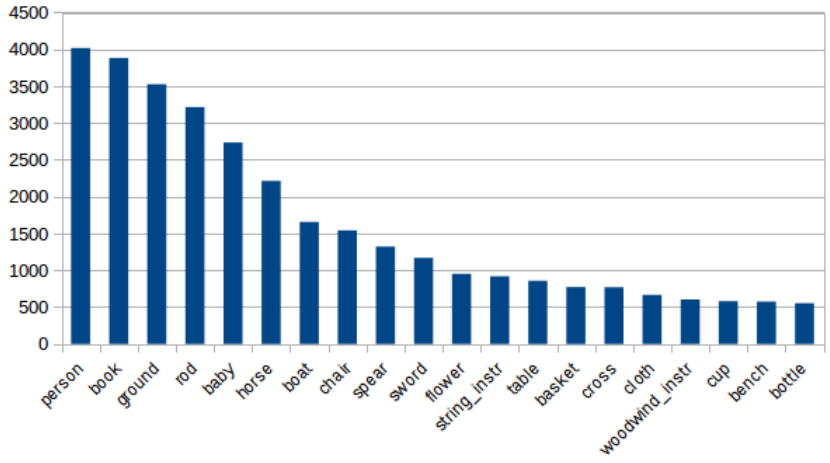


Figure 5.2: Number of total instances per object class in SemArt-HOI dataset.

Figure 5.2 presents a bar chart that shows the total number of instances for each of the 99 object categories in the dataset. This chart illustrates that the most frequent objects in the dataset are person, book, and ground. In addition to objects, the SemArt-HOI dataset contains annotations for verbs and interactions.

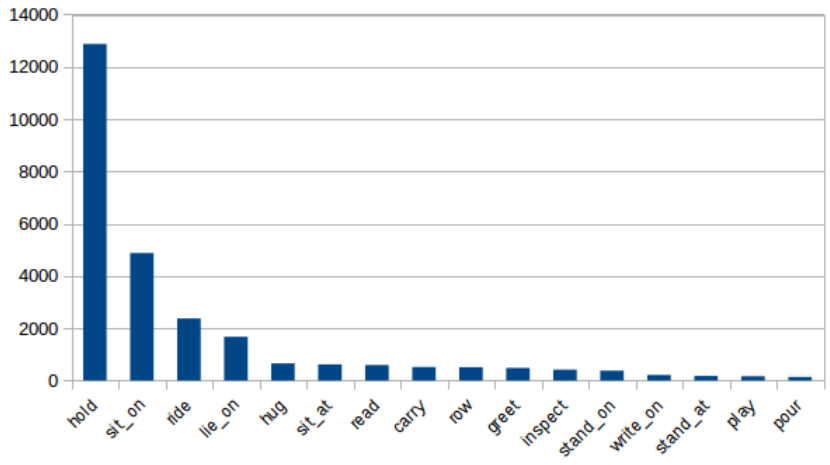


Figure 5.3: Number of total instances per action class in SemArt-HOI dataset.

Figure 5.3 presents a bar chart that shows the total number of instances for each of the 66 verbs in the dataset. The chart illustrates that the most frequent verbs are hold, sit-on, and ride.

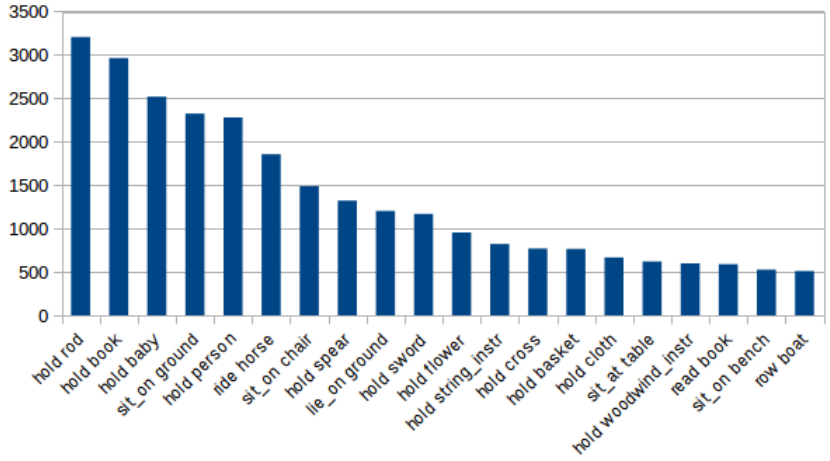


Figure 5.4: Number of total instances per interaction class in SemArt-HOI dataset.

Finally, Figure 5.4 presents a bar chart that shows the total number of instances for each of the 248 interactions in the dataset. This chart illustrates that the most frequent interactions are person-hold-rod, person-hold-book, and person-hold-baby. These charts provide valuable insights into the distribution of objects and interactions in the SemArt-HOI dataset, which can aid in the development and evaluation of HOI detection models.

5.3 Experiments

5.3.1 *Experiments on HICO-DET Dataset*

The models NN1, NN2, and NN3 developed on natural images are trained on the HICO-DET dataset, as there were no painting datasets available yet. During training, a threshold of 0.8 is set for the human detection score and 0.4 for the object detection score to filter out detections. To extract pose features using Regional Multi-person Pose Estimation (RMPE), 17 keypoints are obtained from the human bounding box. These keypoints are then connected with lines of varying gray values (ranging from 0.15 to 0.95) to represent different body parts, constructing the pose map. The size of the final feature vector in both the visual and semantic streams is set to 1x512.

For each detected human, the visual, spatial, and pose features are concatenated together and passed through a fully connected layer, followed by a sigmoid activation function. This process is used to obtain an action prediction score based on the human information. Similarly, another action prediction score is obtained by concatenating the object and union appearance features, followed by a fully connected layer and a sigmoid activation function. These scores are then used to calculate individual cross-entropy losses: $\mathcal{L}cross^H$ for the human and $\mathcal{L}cross^O$ for the object.

The experiments are conducted on the HICO-DET dataset, which is a large dataset for Human Object Interaction (HOI) prediction. The dataset consists of 38,118 training images and 9,658 testing images, covering 80 objects and 117 action verbs. HICO-DET provides annotations for 600 human-object interactions, categorized as Rare or Non-Rare based on their occurrence frequency in the training set. There are 138 Rare and 462 Non-Rare interactions in the dataset.

To extract affordance-based relationships and action verbs, ConceptNet is used as a database. ConceptNet combines data from various sources such as crowd-sourced platforms, expert-created resources, and games with a purpose like Wiktionary and OpenCyc.

The evaluation metric used for performance assessment is the role mean Average Precision (role mAP), following the approach of [58]. A prediction for a human-object interaction is considered correct if the human and object bounding boxes have an Intersection over Union (IoU) greater than 0.5 with the ground-truth boxes, and if the predicted verb class label is accurate.

For human and object detection, a pretrained Faster R-CNN model [101] is utilized. During training, a threshold of 0.8 is set for the human detection score, and 0.4 for the object detection score. These thresholds are determined through experimentation. ResNet-101 [54] is used as the feature extraction backbone. Fine-tuning of the Faster R-CNN model is performed only during testing.

The object nodes in the graph network are represented by their semantic features. BERT [74] is employed as a pretrained model to extract vector representations of words, which have a size of 1x768. To obtain the semantic features of candidate verbs, the sentence composed of the verb and the object is inputted to BERT, and the features of the first word are extracted as the verb features in the context of the

object. Two convolutional layers are applied to the input graph to obtain the final semantic vector representations of the object words and their connected verbs, with a size of 1x512. LeakyReLU activation function with a negative slope of 0.2 is used after each layer of the graph.

The hyperparameters for the total loss are set as follows: λ_1 and λ_2 are both 1, and λ_3 is set to 2. The margin for the cosine loss is set to 0.1. The model is trained using Stochastic Gradient Descent (SGD) for 10 epochs, with a learning rate of 0.001, weight decay of 0.0005, and momentum of 0.9. The experiments are conducted on a 1xV100 NVIDIA 32GB GPU. More than 18 networks are designed, and the performance of the best three models is reported.

The three different models are trained and tested on the HICO-DET dataset under Default and Known Object settings. The mean average precision (mAP) is used as the evaluation metric, following the method of [58]. The mAP results for the proposed models are presented in Table 5.5. From the results, it can be observed that the proposed model significantly improves HOI prediction mAP by more than 12% on the HICO-DET Full test set. However, end-to-end transformer-based model QAHOI [83] still outperform two-stage models on natural images due to the implicit attention mechanism represented in vision transformers for visual contextual representation. In QAHOI, Chen et al. [83] utilize query-based anchors and a multi-scale architecture to extract features from different spatial scales and predict all elements of an HOI instance. The multi-scale architecture in QAHOI allows for capturing the spatial variability of objects in different scales, leading to improved localization and more accurate HOI detection. The use of query-based anchors provides a flexible and adaptable approach, enabling the model to effectively handle the variability in human-object interactions.

In addition to evaluating the proposed models on the HICO-DET dataset, the best performing model, NN3, is further tested on the Watercolor subset from the BAM dataset. It is compared to the state-of-the-art HOI prediction model, QAHOI. The evaluation on the Watercolor subset consists of testing the model on 235 images using three different object detectors: YOLOv3 [137], Faster R-CNN [101], and DETR [81], which is a transformer-based object detector. All of these object

Table 5.5: Performance of the proposed systems compared to the SOTA model on the HICO-DET test sets (%mAP).

Model	Detector	Feature Backbone	HICO-DET					
			Default			Known Object		
			Full	Rare	NoneRare	Full	Rare	NoneRare
Base[33]	COCO	ResNet50	14.7	13.26	15.13	-	-	-
NN1			18.99	14.54	20.32	19.72	16.91	20.56
NN2			20.21	15.69	21.56	21.65	19.84	22.19
NN3			22.73	21.37	23.14	25.86	24.57	26.24
NN3	HICO-DET	ResNet101	27.26	21.92	28.85	29.27	24.77	30.61
QAHOI		Swin-Large	**35.78**	**29.8**	**37.56**	**37.59**	**31.66**	**39.36**

Table 5.6: Performance of NN3 and QAHOI on the Watercolor dataset.

Model	Object Detector	Wrong Object Detected	Correct Object & Wrong Interaction	Correct HOI
NN3	YOLOv3	174/235	31/235	30/235
	Faster-RCNN	**127/235**	49/235	**59/235**
	DETR	169/235	35/235	31/235
	Ground Truth	-	103/235	**132/235**
QAHOI	-	134/235	53/235	48/235

detectors are pretrained on the MS-COCO dataset [25]. The performance of the model is also evaluated using ground truth detections from the Watercolor dataset.

Based on the analysis of the results in Table 5.6, it is clear that the proposed model outperforms the current state-of-the-art system on the Watercolor dataset when tested with Faster R-CNN as the object detector. However, both models still face significant challenges in accurately detecting human-object interactions in the context of paintings.

One of the main challenges is the correct detection of objects in the paintings. More than 50% of the images in the Watercolor dataset did not have a correct object detected, which negatively impacted the performance of both models. Improving the object detection step is crucial for enhancing the overall performance of HOI prediction models on such artistic datasets. Additionally, the proposed model struggled to predict 44% of the actions correctly. This can be attributed to the nature of paintings, where the depicted actions may not always align with real-world scenarios or may require a deeper understanding of the artistic context.

5.3.2 *Experiments on Semart-HOI Dataset*

To evaluate the effectiveness of the proposed systems on the SemArt-HOI dataset, I follow the approach described in [58] and measure their performance using the role mean average precision (role mAP) metric. This metric considers a predicted human-object interaction to be accurate if both the human and object bounding boxes have an intersection over union (IoU) greater than 0.5 with the corresponding ground-truth boxes, and the predicted verb class label for the interaction pair is correct. For detecting humans and objects in the SemArt-HOI dataset, you fine-tune the Faster R-CNN [101] model for 150 epochs. The threshold values of 0.8 for human detection score and 0.4 for object detection score are determined through experimentation. The ResNet-101 [54] architecture is used as a backbone for feature extraction.

In the single-task learning model HOI-Paint, the hyperparameters λ_1, λ_2, λ_3, and λ_4 are set to 1, representing equal importance for all loss components. The models are trained using Stochastic Gradient Descent (SGD) for 10 epochs. The learning rate is set to 0.001, the weight decay is set to 0.0005, and the momentum is set to 0.9. The training process is conducted on a single Nvidia V100 GPU with

Table 5.7: Performance comparison of my models in Single-Task Learning (STL) and Multi-Task Learning (MTL) settings, HOI-Paint and HOI-Paint-MTL, with the SOTA HOI detection model on the SemArt-HOI test sets (%mAP).

Model	mAP	Training-Time	Training-GPU
HOI-Paint	18.32	1 day 18 hours	1xNvidia V100
HOI-Paint-MTL	**18.64**	2 days 1 hour	
QAHOI [83]	17.13	8 days	3xNvidia V100

32 GB memory capacity.

In the multi-task learning model, the hyperparameters are set as $\lambda_1 = \lambda_2 = \lambda_4 = \lambda_5 = \lambda_6 = \lambda_7 = \lambda_8 = 1$ and $\lambda_3 = 0$. Table 5.7 provides a comparison of the performance of HOI-Paint and HOI-Paint-MTL with the state-of-the-art QAHOI model. To ensure a fair comparison, I trained the QAHOI model on the SemArt-HOI dataset for 150 epochs, using 3 parallel 32 GB Nvidia V100 GPUs. The results demonstrate that my proposed MTL system achieves better performance than the STL system, as well as the QAHOI model.

It is interesting to note that the QAHOI model leverages contextual information from the entire image, rather than solely relying on the appearance features of interacting pairs. However, my results suggest that this approach does not improve the accuracy of HOI detection in paintings. Instead, my model's performance indicates that focusing on the appearance features of each individual interacting pair, rather than the context of the entire scene, leads to more accurate predictions.

Moreover, my proposed two-stage models also require significantly less time and computational resources than the state-of-the-art one-stage model, QAHOI. Specifically, my models only requires one GPU for training and can be trained in a shorter amount of time, whereas QAHOI requires three GPUs and a longer training time. The reduced training time and computational requirements of my models allow for faster deployment.

As depicted in Table 5.7, the incorporation of broader semantic context in the proposed systems overcomes the limitations of relying solely on visual cues. This integration allows for a more comprehensive analysis of the painting, resulting in improved effectiveness and accuracy of HOI detection. By considering the semantic context, the systems are able to gain a deeper understanding of the meaning and interpretation of the artwork. This enhancement contributes to a more nuanced understanding of the relationships between humans and objects in the painting, ultimately enriching the overall analysis of the artwork.

HOI-Paint-MTL's success in outperforming the QAHOI model provides further evidence to support the conclusion that incorporating shared visual features in MTL leads to improved performance in HOI detection tasks. The MTL approach allows for shared representations across tasks, enabling the model to leverage visual features that are relevant to multiple tasks simultaneously. This leads to a more comprehensive understanding of the input image, which improves the accuracy of the HOI task, as well as the additional attributes being predicted.

Therefore, I demonstrate that incorporating multi-task learning into my model

training process, with respect to predicting image attributes such as type, school, timeframe, and author, led to a notable improvement in the model's performance for detecting interactions in paintings. By leveraging information from these other tasks, my MTL model was able to enhance the interaction detection capabilities by 0.32% and outperform QAHOI by 1.51%.

Table 5.8: The training loss of HOI when trained in all possible 2TL settings.

Setting		Task Combination	Training Loss of HOI
Trained in	STL	HOI	0.046741
	2TL	HOI+TYPE	0.0322
		HOI+SCHOOL	0.0543
		HOI+TIMEFRAME	0.0481
		HOI+AUTHOR	**0.0143**

Table 5.9: The estimated training loss of HOI when trained in different high order settings (3TL,4TL,5TL) based on HOA.

Setting		Task Combination	HOI Loss using HOA
Trained in	3TL	HOI+TYPE+SCHOOL	0.4325
		HOI+TYPE+TIMEFRAME	0.04015
		HOI+TYPE+AUTHOR	0.02325
		HOI+SCHOOL+TIMEFRAME	0.0512
		HOI+SCHOOL+AUTHOR	0.0343
		HOI+TIMEFRAME+AUTHOR	0.0312
	4TL	HOI+TYPE+SCHOOL+TIMEFRAME	0.04487
		HOI+TYPE+SCHOOL+AUTHOR	0.0336
		HOI+TYPE+TIMEFRAME+AUTHOR	0.03153
		HOI+SCHOOL+TIMEFRAME+AUTHOR	0.0389
	5TL	HOI+TYPE+SCHOOL+TIMEFRAME+AUTHOR	0.037225

Based on my main objective of improving the performance of HOI detection, I apply Higher Order Approximation to my HOI-Paint-MTL model to determine the best tasks to train with HOI for optimal HOI detection. I conducted a 2-Task Learning (2TL) experiment to train HOI with the other four tasks for 10 epochs, using equal weights. Specifically, I trained HOI detection with Type, School, Timeframe, and Author separately, and evaluated the training loss, as shown in Table 5.8.

After obtaining these losses, I estimated the HOI loss when trained in higher order settings (3TL, 4TL, and 5TL) using HOA as shown in Table 5.9. From the results in Tables 5.8 and 5.9, we can see that the best combination for HOI is to be trained with Author in a 2TL setting which leads to the lowest HOI training loss. In Table 5.10, I compared the performance of my model in STL, MTL, and 2TL with Author.

Table 5.10: Performance comparison of HOI-Paint, HOI-Paint-MTL, and HOI-Paint-2TL with author on the SemArt-HOI test sets (%mAP).

Model	mAP
HOI-Paint	18.32
HOI-Paint-MTL	**18.64**
HOI-Paint-2TL (Author)	17.52

However, the results show that the model did not give the highest mean average precision as estimated by HOA. A lower training loss indicates that the model is better able to fit the training data. However, a lower training loss does not necessarily mean that the model will have a higher mAP. This is because the model may be overfitting the training data, which means that it is learning the noise in the data rather than the underlying patterns. A higher mAP indicates that the model is better able to generalize to new data. This is because the mAP is a measure of how well the model is able to predict the labels of the test data. Therefore, the relationship between the training loss and testing mAP can be complex. In some cases, a lower training loss can lead to a higher mAP. However, in other cases, a lower training loss can lead to a lower mAP. This is because the model did overfit the training data.

To gain further insight into the impact of the various tasks learned simultaneously, I conducted additional experiments by training HOI-Paint-MTL with different loss weighting methods. By varying the loss weights for the different tasks, I was able to assess how much each task contributes to the overall performance of the model. This information helps us to better understand the importance of each task and how they relate to one another.

Uncertainty Weighting (UW) is used to balance the contributions of each task to the overall performance of the model. By assigning higher weights to uncertain tasks, the model can focus more on improving its performance on those tasks, while still considering the other tasks. Thus, I use the inverse of the variance of the task-specific loss as the weight for each task; this means that tasks with high variance, indicating high uncertainty, will have higher weights, while tasks with low variance, indicating low uncertainty, will have lower weights. In particular, the weights optimize the model weights W and the noise parameters σ_1,σ_2 to minimize the following objective:

$$\mathcal{L}(W, \sigma_1, \sigma_2) = \frac{1}{2\sigma_1^2}\mathcal{L}_1(W) + \frac{1}{2\sigma_2^2}\mathcal{L}_2(W) + log\sigma_1\sigma_2 \ , \tag{5.1}$$

where, the loss functions $\mathcal{L}_1$, $\mathcal{L}_2$ belong to the first and second task respectively. By minimizing the loss $\mathcal{L}$ with respect to the noise parameters σ_1,σ_2, one can essentially balance the task-specific losses during training.

Random Loss Weighing (RLW) is a technique used that assigns random weights to each task during training. The idea behind this technique is to introduce noise into the training process, which helps the model learn more robust and generalizable representations. The main advantage of random loss weighing is that it avoids the need to manually set the task weights, which is challenging, especially when the

tasks have different scales or levels of difficulty. By randomly assigning the weights, the model is forced to learn to adapt to the different tasks without being biased towards any particular task. Thus to apply RLW in my system, I use a Gaussian distribution to sample the weights, which results in smooth updates to the task weights over time.

Dynamic Weight Average (DWA) computes a weighted average of the model parameters, where the weights are determined dynamically based on the recent history of the model's performance. The main advantage of DWA is that it allows the model to adapt to changes in the input distribution, which can occur frequently in online learning scenarios. By adjusting the weights based on recent performance, DWA gives more weight to the most recent and accurate model parameters, while discounting older and potentially less relevant parameters. DWA is typically used in conjunction with stochastic gradient descent (SGD) or other optimization algorithms, where the model parameters are updated in small batches based on the gradient of the loss function. Instead of using a fixed learning rate, DWA adjusts the weights of the model parameters based on a running average of the loss function over a recent window of training examples. DWA is implemented my model by using an Exponential Moving Average (EMA) to compute the dynamic weights. The EMA gives more weight to more recent loss values, and less weight to older values, so that the weights are updated gradually and smoothly over time. In DWA, the task-specific weight ω_i for task i at step t is set as:

$$\omega_i(t) = \frac{Nexp(r_i(t-1)/T)}{\sum_n exp(r_n(t-1)/T)} \ , r_n(t-1) = \frac{L_n(t-1)}{L_n(t-2)} \ , \tag{5.2}$$

where, N is the number of tasks. The scalars r_n $(\cdot)$ estimate the relative descending rate of the task-specific loss values L_n. The temperature T controls the softness of the task weighting. When the loss of a task decreases at a slower rate compared to other tasks, the task-specific weight in the loss is increased. Thus, the target is to minimize the total loss $\mathcal{L}$ defined as:

$$\mathcal{L} = \sum \omega_i \mathcal{L}_i \ , \tag{5.3}$$

In Scenario1 and Scenario2, I manually set the loss weights for the different tasks. For Scenario1, the Type, School, Time and Author losses are given a weight equal to 0.15, while the HOI loss weight is set to 0.4. In Scenario2, the Type, School, Time and Author losses are given a weight equal to 0.1, while the HOI loss weight is set to 0.6.

Table 5.11 presents the results of various experiments conducted using different loss weighting strategies. The Random Loss Weighting strategy demonstrated the highest mAP, resulting in an overall performance improvement of 0.07%. RWL helps the model learn to adapt to all tasks without being biased towards any one specific task, which leads to a more balanced optimization process and improved performance on all tasks. The random weighting also encourages the model to learn the relationships between tasks, which can further improve its ability to generalize to new data.

Table 5.11: Performance comparison of my models using different loss weighting methods on the SemArt-HOI test sets (%mAP).

Method	mAP
HOI-Paint	18.32
HOI-Paint-MTL-EW	18.64
HOI-Paint-MTL-UW	18.71
HOI-Paint-MTL-RLW	**18.78**
HOI-Paint-MTL-DWA	17.63
HOI-Paint-MTL-Scenario1	18.59
HOI-Paint-MTL-Scenario2	18.35

5.3.3 *Ablation Studies*

To evaluate the influence of each representation obtained from the episodic memory, I train the HOI-Paint model without the semantic memory and vary the weights of loss associate to each pf the human, object, and ROI representations. Table 5.12 reports the evaluation results of my STL model with episodic memory stream only, HOI-Paint-Episodic, on the SemArt-HOI dataset with different loss weights. The weights were adjusted according to the episodic loss function. At first, I gave equal weights to the human, object, and region of interest (ROI) losses. However, I observed that the model's performance improved after removing the ROI loss. This is because the ROI loss was found to be less informative in the context of my dataset, as the objects and humans in paintings are often tightly cropped and centered. Removing the ROI loss allowed the model to focus more on the appearance features of the interacting pairs, leading to better performance.

To further analyze the influence of the human and object losses on the overall performance, I experimented with different weights for each loss. I assigned the same weight to the human and object losses and varied their values from 0.1 to 1.0 with a step size of 0.1. The ROI loss was excluded from these experiments. Interestingly, I found that the model achieved the highest mean average precision (mAP) when the human and object losses were equally weighted. This suggests that the contributions of the human and object losses to the total loss are roughly equivalent, and that the model benefits from learning both equally.

Overall, as shown in Table 5.12, my experiments demonstrate that careful selec-

Table 5.12: Performance of the proposed HOI-Paint-Episodic system on the SemArt-HOI test sets (%mAP) under different loss weighing settings.

Model	Setting	mAP
HOI-Paint-Episodic	$\lambda_1 = \lambda_2 = \lambda_3 = 1$	17.54
	$\lambda_1 = \lambda_2 = 1, \lambda_3 = 0$	**17.93**
	$\lambda_1 = 0.3, \lambda_2 = 0.7, \lambda_3 = 0$	17.76
	$\lambda_1 = 0.7, \lambda_2 = 0.3, \lambda_3 = 0$	17.73
HOI-Paint		**18.32**

tion and weighting of the loss terms significantly impacts the performance of the HOI detection model. By removing less informative loss terms and tuning the weights of the remaining ones, I was able to improve the performance of episodic memory module by 0.39%. However, the absence of the semantic stream caused a decrease in the HOI mAP by 0.39%. This shows that the integration of the semantic context enables a more comprehensive analysis of the painting, leading to enhanced effectiveness and accuracy in detecting HOIs. Thus, the inclusion of semantic context allows the systems to attain a deeper understanding of the artwork's meaning and interpretation.

Table 5.13: Performance of the proposed STL system in HOI-Paint on the SemArt-HOI test sets (%mAP) with and without spatial features.

Model	mAP
HOI-Paint	**18.32**
HOI-Paint-Episodic	17.93
HOI-Paint-Episodic-w/o-Spatial	13.77

To evaluate the impact of the spatial stream in the model, I trained a version of the HOI-Paint-Episodic model using only visual features. The results in Table 5.13 demonstrate that incorporating spatial features improves the model's performance by more than 4%. While it is true that physical or geometric relations may not always be applicable in paintings, this dataset primarily consists of images depicting portraits, religious icons, battle scenes, sculptures, and interior monuments where such relations are likely to be present.

To further investigate the effect of the inclusion of the knowledge graph in the model, Table 5.14 presents the HOI mAP on the different types of images. We can see that the model performs worse in historical paintings and best in portraits. When it comes to historical paintings, common sense does not always apply or be reliable for several reasons such as historical context where historical paintings depict events, settings, and people from different time periods and cultural contexts. The societal norms, customs, and beliefs of those historical periods differ significantly from our present-day common sense. Actions, gestures, or symbols portrayed in historical

Table 5.14: Performance of my HOI-Paint-Episodic and HOI-Paint models on the some types of paintings in the test SemArt-HOI dataset (%mAP)

Type	Number of test images	HOI-Paint-Episodic	HOI-Paint
Portrait	476	19.91	**21.44**
Religious	3787	17.53	18.81
Interior	259	17.82	18.67
Mythological	940	17.64	17.72
Landscape	1038	15.65	15.75
Genre	858	14.51	14.61
Historical	418	13.93	**13.33**

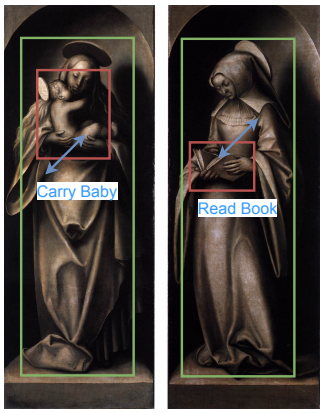
(a) Historical painting.

(b) Portrait painting.

Figure 5.5: Two types of paintings from the SemArt-HOI dataset. In figure 5.5a, semantics do not apply, whereas in figure 5.5b semantics apply.

paintings have had different meanings or connotations that are no longer apparent to contemporary viewers. Moreover, historical paintings frequently employ symbolism and allegory to convey complex ideas or represent abstract concepts. These symbolic elements do not align with everyday common sense interpretations. Artists use visual metaphors, religious or mythological references, or specific iconography that require contextual knowledge or historical understanding to interpret correctly. In addition, artists often exercise creative license in historical paintings, emphasizing artistic expression over strict adherence to factual accuracy. They distort proportions, exaggerate certain features, or employ dramatic lighting and composition for aesthetic or narrative purposes. These artistic choices can deviate from common sense expectations and realism, intentionally distorting or stylizing historical events or figures. It is essential to approach historical paintings with an awareness of their specific historical, cultural, and artistic contexts. The interpretation of these artworks requires a combination of historical knowledge, contextual understanding, and an appreciation for the artistic intentions and techniques employed by the artists.

While common sense can play a role in analyzing historical paintings, it should be complemented by a deeper exploration of the historical period and the artistic conventions prevalent at the time of their creation. However, common sense often applies in portrait paintings because they typically aim to depict the physical appearance and character of the subject in a realistic manner. For instance, portrait paintings strive to capture the likeness and physical features of the individual being portrayed. Common sense enables viewers to recognize familiar facial expressions, body proportions, and other visual cues that are consistent with their everyday experiences of observing and interacting with people. While common sense can provide a useful framework for interpreting portrait paintings, it's important to note that artists also have the ability to manipulate and stylize their subjects. They may emphasize certain features, exaggerate expressions, or incorporate symbolic elements to convey deeper meaning or evoke specific emotions. Therefore, while common sense



(a) Lack of colors, and context.

(b) Lack of scale.

(c) Lack of scale and context.

Figure 5.6: HOI detections on SemArt-HOI dataset.

serves as a starting point, a nuanced analysis of the artist's intention, cultural context, and artistic techniques is crucial for a comprehensive understanding of portrait paintings (Figure 5.5).

Paintings have their own unique characteristics and may depict spatial relationships and semantic features that differ from those in natural images. Retraining the model with these specialized spatial and semantic streams allows it to learn new spatial relationships that are specific to paintings and capture semantic features that are relevant in the context of paintings. This adaptation enhances the model's ability to understand and predict human-object interactions in the context of artwork, where common-sense or physical relationships may differ from those in natural images. Therefore, the addition of the spatial and semantic streams and retraining the model specifically for paintings improves the model's performance by accounting for the unique spatial relationships and semantic features present in artwork, leading to enhanced HOI predictions.

Figure 5.6 presents qualitative results of my proposed model on a few examples of such images. For instance, in Figure 5.6a, the image contains only a black and white sketch of a human holding a book. Despite the lack of color and contextual information, my model can accurately detect the interaction between the human and the book. This demonstrates the robustness of my model to handle images with missing information.

Similarly, in Figure 5.6b, my model can detect the interaction between a human and a horse in an image that lacks contextual information. This further highlights the capability of my model to accurately detect HOIs even in challenging scenarios.

Moreover, in Figure 5.6c, my model was able to detect the interactions in a scene where context and scale lack. This showcases the effectiveness of my model in handling diverse and complex scenarios in painting analysis. Overall, the qualitative results provide evidence for the effectiveness and robustness of my proposed model for HOI detection in paintings.

Chapter 6

Discussion

My proposed approach for HOI detection in paintings differs from the QAHOI model in terms of reliance on visual contextual information. While QAHOI utilizes visual context from the entire image, my approach focuses specifically on individual interacting pairs. The results demonstrate that contextual information from the entire image is not as useful for HOI detection in paintings. This finding is significant because contextual information plays a critical role in other applications, such as object detection in natural images. However, in the case of paintings, where the composition is often highly stylized and structured, concentrating on individual interacting pairs can provide more informative insights compared to considering the overall scene context. Specifically, my research reveals that incorporating visual features from the region encompassing the human and the object can actually decrease the performance of the model. This finding highlights that, in paintings, visual context can have a negative impact on interaction prediction due to the unique nature of paintings and their presentation style.

Moreover, my proposed model takes into account the specific requirements of paintings, such as the need for contextual knowledge and semantic relationships. Incorporating semantic context through the Graph Convolution Network (GCN) is an essential aspect of the proposed system. While visual context is not always present or reliable in paintings, semantic context provides valuable insights and overcomes the limitations of relying solely on visual cues. In some cases, visual context can be misleading or ambiguous, making it challenging to accurately infer human-object interactions. By leveraging the GCN, the model integrates semantic relationships and contextual knowledge into the HOI detection process. The knowledge graph formed by the GCN captures the connections between different elements, such as verbs and object categories, and represents the underlying semantic context of the painting. This semantic context enriches the model's understanding of the artwork and facilitates more accurate inference of human-object interactions.

To assess the impact of the lack of visual context on HOI prediction, the proposed HOI-Paint model and QAHOI were evaluated on the Watercolor dataset. The results were compared using Figures 6.1 and 6.2, which showcase the HOI detection output of both models on paintings where context is absent and the principles of physics do not apply. The comparison reveals that the proposed STL model out-

Figure 6.1: HOI detection of HOI-Paint model on the Watercolor dataset. Left to right: hold baby, ride cart, hold person

Figure 6.2: HOI detection of QAHOI model on the Watercolor dataset. Left to right: sit_on bench, hold cloth, hold book

performed QAHOI in predicting the correct interactions in these contextually challenging paintings. QAHOI heavily relies on visual contextual features that are not present or discernible in these images, leading to inaccurate predictions. In contrast, the MTL model demonstrated its capability to extract discriminative features from each detected instance without relying on contextual information. By disregarding the context of the painting and focusing on the individual instances, my model effectively captured and represented essential visual patterns for HOI detection. This ability to extract discriminative features allowed the model to overcome the limitations imposed by the absence of context and the violation of physics principles in these paintings. As a result, the proposed STL model achieved more accurate and reliable HOI predictions compared to QAHOI in this challenging scenario.

Human pose features were not included in the proposed models because in paintings, the depiction of human pose is often stylized and may not accurately represent the actual physical poses. The artistic interpretation and style employed in paintings can lead to exaggerated or abstract representations of human figures, making it challenging to extract reliable pose information as shown in Figure 6.3. There-

fore, incorporating pose features may not contribute significantly to the accuracy of HOI detection in the context of paintings. Instead, my focus was on capturing visual patterns, semantic relationships, and contextual information to enhance the detection of human-object interactions in paintings.

Figure 6.3: Image from SemArt-HOI dataset showing the complex human poses styled in the painting.

Moreover, paintings have unique characteristics and challenges that differ from natural images. They often lack rich visual cues and contextual information, making the detection of human-object interactions more complex. My HOI-Paint-MTL model is specifically designed to address these challenges by incorporating additional classification tasks and leveraging shared representations across multiple tasks. This allows the model to gain a deeper understanding of the painting and its elements, enhancing the accuracy of HOI detection. This comprehensive understanding of the input image improves the accuracy of the HOI task, as well as the additional attributes being predicted. The experimental results provide evidence to support the conclusion that incorporating shared visual features in MTL leads to improved performance in HOI detection tasks. Specifically, I demonstrate that incorporating MTL into my model training process to predict image attributes such as type, school, timeframe, and author leads to a notable improvement in the model's performance for detecting interactions in paintings. By leveraging information from these other tasks, the model was able to enhance its interaction detection capabilities, and outperform the state-of-the-art QAHOI model. This improvement highlights the

Table 6.1: Performance comparison of my model in Single-Task Learning (STL) and Multi-Task Learning (MTL) settings, using different object detectors, on the SemArt_HOI test sets (%mAP).

Model	Object Detector	mAP_HOI
HOI-Paint	FASTER-RCNN	17.93
	DETR	17.22
	GROUND TRUTH	48.83
HOI-Paint-MTL	FASTER-RCNN	18.64
	DETR	17.56
	GROUND TRUTH	56.15

benefits of using a multi-task approach for HOI detection in paintings. The results demonstrate that shared visual features improves the accuracy and efficiency of the model, leading to better performance in detecting interactions. These findings have important implications for the development of HOI detection models and may have broader applications in computer vision.

To evaluate the significance of the object detector in this process, I compared the performance of my model with that of Faster R-CNN, DETR, and ground truth detections. The results, as shown in Table 6.1, indicate that the model's performance is optimal when tested with ground truth object detections followed by Faster-RCNN and than DETR. These findings imply that object detection in paintings differs from that in natural images, where contextual information from the encoder can enhance the model's detection ability. In addition, since there are many small or intricately shaped objects in paintings, which may lack contextual information, CNN-based models outperform transformer-based models. This suggests that the presence or absence of contextual information should be taken into account while selecting the object detection approach in paintings.

In addition to better performance, the proposed two-stage model is more efficient in terms of time and computational resources than the state-of-the-art one-stage model, QAHOI. My model requires significantly less time and computational resources than QAHOI. The proposed model only requires one GPU for training, whereas QAHOI requires three GPUs. This indicates that the proposed model is less computationally demanding and therefore, more cost-effective than QAHOI. Moreover, my model is trained in a shorter amount of time, further emphasizing the efficiency of their approach. The reduced training time and computational requirements of the proposed model have important practical implications. Specifically, they allow for faster deployment of the model, which is crucial in many applications,, where decisions need to be made quickly and accurately, faster deployment of models can have a significant impact on the overall performance of the system. Thus, the proposed two-stage model is better than the state-of-the-art one-stage model, QAHOI, not only in terms of precision but also in terms of time and computational resources. The reduced training time and computational requirements of the proposed model can lead to more efficient and cost-effective solutions, and enable faster deployment in real-world applications.

Chapter 7

Conclusion and Future Work

The main objective of this research is to develop a specialized system for detecting Human-Object Interactions in paintings, which hold important cultural value. In contrast to existing frameworks, my proposed model focuses specifically on the appearance features of each individual interacting pair depicted in the artwork. Additionally, the incorporation of semantic features proves highly advantageous in enhancing the model's performance, as they contribute to improved accuracy.

Through experimental evaluations, I demonstrated that my model outperforms the state-of-the-art HOI detection model designed for natural images, which heavily relies on visual context representations. This finding suggests that visual context is misleading in the context of HOI detection in paintings. Furthermore, I found that removing the Region of Interest features, which represent the contextual information of the interaction between the human and object, actually leads to improved performance in my model. This indicates that the surrounding context in paintings does not provide significant predictive value for HOI detection. When considering the human and object equally without incorporating the surrounding context, my model's predictive ability remained unaffected, which suggests that both human and object features are equally informative for HOI prediction in paintings.

Additionally, the multi-task learning approach employed in my two-stage model, contributes to its superiority. By introducing four supplementary classification tasks, the model benefits from complementary information that improves the overall HOI prediction. The shared representations learned from these tasks enhance the model's understanding of the painting and its elements. Moreover, the inclusion of multiple artistic tasks in the multi-task learning framework facilitates a more comprehensive analysis of the painting. Each task provides complementary information about the artwork, such as the artist, timeframe, type, or author, which contributes to a deeper understanding of the painting's context. By considering these additional tasks, the model gains a more holistic perspective and can capture a wider range of features and patterns, ultimately improving the accuracy of HOI detection. Multi-task learning facilitates knowledge transfer between related tasks, where the shared representations learned across tasks can help to transfer knowledge and leverage the learned information from one task to benefit the others.

Furthermore, the Random Loss Weighing (RLW) strategy proved to be the most

effective in my model. In RLW, different tasks are assigned random weights, encouraging the model to learn robust and adaptable representations across tasks. By assigning random weights, the model discovers interactions and dependencies between tasks that might not be apparent with fixed or pre-defined weights. Regularization in RLW helps prevent overfitting and improves the model's generalization ability across tasks. Furthermore, RLW helps prevent the dominance of a single task and encourages the model to allocate resources more evenly among different tasks. This approach enhances the overall performance and flexibility of the multi-task learning framework.

In the HOI-Paint STL model, multiple losses are jointly optimized during training. However, a potential future direction is to explore loss weighting strategies to determine the optimal weights for each loss that contributes to the HOI prediction in HOI-Paint. By assigning appropriate weights to different losses, the model can effectively balance their impact on the overall prediction performance. This approach would allow for a more fine-grained control over the learning process and potentially improve the model's accuracy and generalization ability. Further investigation into loss weighting strategies can lead to enhancements in the training and optimization of the HOI-Paint for HOI detection in paintings.

Given the wide variation in paintings based on aspects such as genre and style, developing a robust HOI detection model that considers these factors can be valuable. By incorporating information about the genre or style of the painting, the model can adapt and adjust the weights of each stream accordingly. This approach allows the model to account for the specific characteristics and nuances associated with different genres or styles, ultimately improving the accuracy and reliability of HOI detection.

Moreover, the current SemArt-HOI dataset includes only action labels for interacting pairs, which means that pairs with no interaction are not labeled at all. This presents a limitation in the dataset, as it does not allow for the distinction between interacting and non-interacting pairs. To address this, adding a 'no_interaction' action label can improve the dataset and make it more comprehensive. In addition, the dataset can be further be improved by allowing for multiple action labels to be assigned to the same interacting pair. By applying a multi-class interaction classification, the model can accurately classify such interactions and provide more detailed information about the nature of the interaction. Overall, improving the dataset in these ways can lead to more accurate and comprehensive HOI detection models.

Bibliography

[1] R. D. Beer, "Dynamical approaches to cognitive science," *Trends in cognitive sciences*, vol. 4, no. 3, pp. 91–99, 2000.

[2] J. R. Anderson, *Learning and memory: An integrated approach*. John Wiley & Sons Inc, 2000.

[3] S. Smirnov and A. Eguizabal, "Deep learning for object detection in fine-art paintings," in *2018 Metrology for Archaeology and Cultural Heritage (MetroArchaeo)*, IEEE, 2018, pp. 45–49.

[4] H.-J. Jeon, S. Jung, Y.-S. Choi, J. W. Kim, and J. S. Kim, "Object detection in artworks using data augmentation," in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, IEEE, 2020, pp. 1312–1314.

[5] D. Kadish, S. Risi, and A. S. Løvlie, "Improving object detection in art images using only style transfer," in *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2021, pp. 1–8.

[6] M.-C. Marinescu, A. Reshetnikov, and J. M. López, "Improving object detection in paintings based on time contexts," in *2020 International Conference on Data Mining Workshops (ICDMW)*, IEEE, 2020, pp. 926–932.

[7] S. Bianco, D. Mazzini, P. Napoletano, and R. Schettini, "Multitask painting categorization by deep multibranch neural network," *Expert Systems with Applications*, vol. 135, pp. 90–101, 2019.

[8] J. Zujovic, L. Gandy, S. Friedman, B. Pardo, and T. N. Pappas, "Classifying paintings by artistic genre: An analysis of features & classifiers," in *2009 IEEE International Workshop on Multimedia Signal Processing*, IEEE, 2009, pp. 1–5.

[9] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.

[10] M. Čuljak, B. Mikuš, K. Jež, and S. Hadjić, "Classification of art paintings by genre," in *2011 Proceedings of the 34th International Convention MIPRO*, IEEE, 2011, pp. 1634–1639.

[11] S. Agarwal, H. Karnick, N. Pant, and U. Patel, "Genre and style based painting classification," in *2015 IEEE Winter Conference on Applications of Computer Vision*, IEEE, 2015, pp. 588–594.

[12] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, Ieee, vol. 2, 1999, pp. 1150–1157.

[13] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, pp. 145–175, 2001.

[14] W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition," in *International workshop on automatic face and gesture recognition*, Citeseer, vol. 12, 1995, pp. 296–301.

[15] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[16] S.-G. Lee and E.-Y. Cha, "Style classification and visualization of art painting's genre using self-organizing maps," *Human-centric Computing and Information Sciences*, vol. 6, no. 1, pp. 1–11, 2016.

[17] B. Saleh and A. Elgammal, "Large-scale classification of fine-art paintings: Learning the right metric on the right feature," *arXiv preprint arXiv:1505.00855*, 2015.

[18] E. Cetinic and S. Grgic, "Genre classification of paintings," in *2016 international symposium elmar*, IEEE, 2016, pp. 201–204.

[19] X. Huang, S.-h. Zhong, and Z. Xiao, "Fine-art painting classification via two-channel deep residual network," in *Pacific RIM conference on multimedia*, Springer, 2017, pp. 79–88.

[20] M. K. Hosain, T. B. Taher, M. M. Rahman, *et al.*, "Genre recognition of artworks using convolutional neural network," in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2020, pp. 1–5.

[21] Q. Wang and G. Feng, "Image style recognition using graph network and perception layer," in *CAAI International Conference on Artificial Intelligence*, Springer, 2021, pp. 565–574.

[22] L. A. Iliadis, S. Nikolaidis, P. Sarigiannidis, S. Wan, and S. K. Goudos, "Artwork style recognition using vision transformers and mlp mixer," *Technologies*, vol. 10, no. 1, p. 2, 2021.

[23] B. L. Menai and M. C. Babahenini, "Recognizing the style of a fine-art painting with efficientnet and transfer learning," in *2022 7th International Conference on Image and Signal Processing and their Applications (ISPA)*, IEEE, 2022, pp. 1–6.

[24] W. Zhao, W. Jiang, and X. Qiu, "Big transfer learning for fine art classification," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[25] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *European conference on computer vision*, Springer, 2014, pp. 740–755.

[26] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "Hico: A benchmark for recognizing human-object interactions in images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1017–1025.

[27] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8359–8367.

[28] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 401–417.

[29] C. Gao, Y. Zou, and J.-B. Huang, "Ican: Instance-centric attention network for human-object interaction detection," *arXiv preprint arXiv:1808.10437*, 2018.

[30] X. Wang, Y. Ye, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6857–6866.

[31] A. Bansal, S. S. Rambhatla, A. Shrivastava, and R. Chellappa, "Detecting human-object interactions via functional generalization.," in *AAAI*, 2020, pp. 10 460–10 469.

[32] Y.-L. Li, S. Zhou, X. Huang, *et al.*, "Transferable interactiveness knowledge for human-object interaction detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3585–3594.

[33] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli, "Learning to detect human-object interactions with knowledge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[34] P. Zhou and M. Chi, "Relation parsing neural network for human-object interaction detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 843–851.

[35] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware multi-level feature network for human object interaction detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9469–9478.

[36] C. Gao, J. Xu, Y. Zou, and J.-B. Huang, "Drg: Dual relation graph for human-object interaction detection," in *European Conference on Computer Vision*, Springer, 2020, pp. 696–712.

[37] Z. Hou, X. Peng, Y. Qiao, and D. Tao, "Visual compositional learning for human-object interaction detection," *arXiv preprint arXiv:2007.12407*, 2020.

[38] O. Ulutan, A. Iftekhar, and B. S. Manjunath, "Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 617–13 626.

[39] Y.-L. Li, X. Liu, H. Lu, *et al.*, "Detailed 2d-3d joint representation for human-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 166–10 175.

[40] X. Zhong, C. Ding, X. Qu, and D. Tao, "Polysemy deciphering network for robust human–object interaction detection," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1910–1929, 2021.

[41] F. Z. Zhang, D. Campbell, and S. Gould, "Spatially conditioned graphs for detecting human-object interactions," *arXiv preprint arXiv:2012.06060*, 2020.

[42] Y. Liu, J. Yuan, and C. W. Chen, "Consnet: Learning consistency graph for zero-shot human-object interaction detection," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4235–4243.

[43] Y.-L. Li, X. Liu, X. Wu, Y. Li, and C. Lu, "Hoi analysis: Integrating and decomposing human-object interaction," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[44] Z. Liang, Y. Guan, and J. Rojas, "Visual-semantic graph attention network for human-object interaction detection," *arXiv preprint arXiv:2001.02302*, 2020.

[45] Z. Hou, B. Yu, Y. Qiao, X. Peng, and D. Tao, "Detecting human-object interaction via fabricated compositional learning," *arXiv preprint arXiv:2103.08214*, 2021.

[46] B. Kim, J. Lee, J. Kang, E.-S. Kim, and H. J. Kim, "Hotr: End-to-end human-object interaction detection with transformers," *arXiv preprint arXiv:2104.13682*, 2021.

[47] M. Chen, Y. Liao, S. Liu, Z. Chen, F. Wang, and C. Qian, "Reformulating hoi detection as adaptive set prediction," *arXiv preprint arXiv:2103.05983*, 2021.

[48] C. Zou, B. Wang, Y. Hu, *et al.*, "End-to-end human object interaction detection with hoi transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 825–11 834.

[49] M. Tamura, H. Ohashi, and T. Yoshinaga, "Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information," *arXiv preprint arXiv:2103.05399*, 2021.

[50] A. Zhang, Y. Liao, S. Liu, *et al.*, "Mining the benefits of two-stage and one-stage hoi detection," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[51] D.-J. Kim, X. Sun, J. Choi, S. Lin, and I. S. Kweon, "Detecting human-object interactions with action co-occurrence priors," *arXiv preprint arXiv:2007.08728*, 2020.

[52] W. Liu, D. Anguelov, D. Erhan, *et al.*, "Ssd: Single shot multibox detector," in *European conference on computer vision*, Springer, 2016, pp. 21–37.

[53] B. Kim, T. Choi, J. Kang, and H. J. Kim, "Uniondet: Union-level detector towards real-time human-object interaction detection," in *European Conference on Computer Vision*, Springer, 2020, pp. 498–514.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[55] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[56] J. Dai, H. Qi, Y. Xiong, *et al.*, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[57] B. Zhuang, L. Liu, C. Shen, and I. Reid, "Towards context-aware interaction recognition for visual relationship detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 589–598.

[58] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *2018 ieee winter conference on applications of computer vision (wacv)*, IEEE, 2018, pp. 381–389.

[59] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He, *Detectron*, 2018.

[60] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[61] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2334–2343.

[62] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 863–10 872.

[63] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7103–7112.

[64] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[65] G. Pavlakos, V. Choutas, N. Ghorbani, *et al.*, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 975–10 985.

[66] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *European conference on computer vision*, Springer, 2016, pp. 852–869.

[67] P. Anderson, X. He, C. Buehler, *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[68] Y.-L. Li, L. Xu, X. Liu, *et al.*, "Pastanet: Toward human activity knowledge engine," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 382–391.

[69] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[70] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[71] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

[72] M. E. Peters, M. Neumann, M. Iyyer, *et al.*, *Deep contextualized word representations*, 2018. arXiv: 1802.05365 [cs.CL].

[73] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[74] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[75] H. Wang, W.-s. Zheng, and L. Yingbiao, "Contextual heterogeneous graph network for human-object interaction detection," in *European Conference on Computer Vision*, Springer, 2020, pp. 248–264.

[76] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[77] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng, "Ppdm: Parallel point detection and matching for real-time human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 482–490.

[78] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning human-object interaction detection using interaction points," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4116–4125.

[79] X. Zhong, X. Qu, C. Ding, and D. Tao, "Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection," *arXiv preprint arXiv:2104.05269*, 2021.

[80] H.-S. Fang, Y. Xie, D. Shao, and C. Lu, "Dirv: Dense interaction region voting for end-to-end human-object interaction detection," *arXiv preprint arXiv:2010.01005*, 2020.

[81] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, Springer, 2020, pp. 213–229.

[82] D. Zhou, Z. Liu, J. Wang, *et al.*, "Human-object interaction detection via disentangled transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 568–19 577.

[83] J. Chen and K. Yanai, "Qahoi: Query-based anchors for human-object interaction detection," *arXiv preprint arXiv:2112.08647*, 2021.

[84] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[85] L. Dong, Z. Li, K. Xu, *et al.*, "Category-aware transformer network for better human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 538–19 547.

[86] F. Baldassarre, K. Smith, J. Sullivan, and H. Azizpour, "Explanation-based weakly-supervised learning of visual relations with graph networks," *arXiv preprint arXiv:2006.09562*, 2020.

[87] Y. Song, W. Li, L. Zhang, *et al.*, "Novel human-object interaction detection via adversarial domain generalization," *arXiv preprint arXiv:2005.11406*, 2020.

[88] D. Kim, G. Lee, J. Jeong, and N. Kwak, "Tell me what they're holding: Weakly-supervised object detection with transferable knowledge from human-object interaction.," in *AAAI*, 2020, pp. 11 246–11 253.

[89] T. Zhou, W. Wang, S. Qi, H. Ling, and J. Shen, "Cascaded human-object interaction recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4263–4272.

[90] X. Sun, X. Hu, T. Ren, and G. Wu, "Human object interaction detection via multi-level conditioned network," in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 2020, pp. 26–34.

[91] S. Wang, K.-H. Yap, J. Yuan, and Y.-P. Tan, "Discovering human interactions with novel objects via zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 652–11 661.

[92] Z. Hou, B. Yu, Y. Qiao, X. Peng, and D. Tao, "Affordance transfer learning for human-object interaction detection," *arXiv preprint arXiv:2104.02867*, 2021.

[93] M. Moscovitch, R. Cabeza, G. Winocur, and L. Nadel, "Episodic memory and beyond: The hippocampus and neocortex in transformation," *Annual review of psychology*, vol. 67, pp. 105–134, 2016.

[94] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, IEEE, vol. 1, 2003, pp. I–253.

[95] R. W. Fleming, "Visual perception of materials and their properties," *Vision research*, vol. 94, pp. 62–75, 2014.

[96] M. Bar, "Visual objects in context," *Nature Reviews Neuroscience*, vol. 5, no. 8, pp. 617–629, 2004.

[97] A. Bansal, S. S. Rambhatla, A. Shrivastava, and R. Chellappa, "Spatial priming for detecting human-object interactions," *arXiv preprint arXiv:2004.04851*, 2020.

[98] L. L. Jacoby and M. Dallas, "On the relationship between autobiographical memory and perceptual learning.," *Journal of Experimental Psychology: General*, vol. 110, no. 3, p. 306, 1981.

[99] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[100] A. Salvador, N. Hynes, Y. Aytar, *et al.*, "Learning cross-modal embeddings for cooking recipes and food images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3020–3028.

[101] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[102] K. Nelissen, G. Luppino, W. Vanduffel, G. Rizzolatti, and G. A. Orban, "Observing others: Multiple action representation in the frontal lobe," *Science*, vol. 310, no. 5746, pp. 332–336, 2005.

[103] V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti, "Action recognition in the premotor cortex," *Brain*, vol. 119, no. 2, pp. 593–609, 1996.

[104] D. Bub and M. Masson, "Gestural knowledge evoked by objects as part of conceptual representations," *Aphasiology*, vol. 20, no. 9, pp. 1112–1124, 2006.

[105] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," *arXiv preprint arXiv:1612.03975*, 2016.

[106] E. B. Goldstein, *Cognitive psychology: Connecting mind, research and everyday experience*. Cengage Learning, 2014.

[107] R. A. Poldrack and T. Yarkoni, "From brain maps to cognitive ontologies: Informatics and the search for mental structure," *Annual review of psychology*, vol. 67, pp. 587–612, 2016.

[108] E. K. Miller and J. D. Cohen, "An integrative theory of prefrontal cortex function," *Annual review of neuroscience*, vol. 24, no. 1, pp. 167–202, 2001.

[109] P. W. Burgess, "Strategy application disorder: The role of the frontal lobes in human multitasking," *Psychological research*, vol. 63, no. 3-4, pp. 279–288, 2000.

[110] K. Sakai and R. E. Passingham, "Prefrontal interactions reflect future task operations," *Nature neuroscience*, vol. 6, no. 1, pp. 75–81, 2003.

[111] M. D. Lieberman, "Social cognitive neuroscience: A review of core processes," *Annu. Rev. Psychol.*, vol. 58, pp. 259–289, 2007.

[112] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[113] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[114] G. Strezoski and M. Worring, "Omniart: Multi-task deep learning for artistic data analysis," *arXiv preprint arXiv:1708.00684*, 2017.

[115] A. Belhi, A. Bouras, and S. Foufou, "Leveraging known data for missing label prediction in cultural heritage context," *Applied Sciences*, vol. 8, no. 10, p. 1768, 2018.

[116] N. Garcia, B. Renoust, and Y. Nakashima, "Contextnet: Representation and exploration for painting classification and retrieval in context," *International Journal of Multimedia Information Retrieval*, vol. 9, no. 1, pp. 17–30, 2020.

[117] W. Zhao, D. Zhou, X. Qiu, and W. Jiang, "How to represent paintings: A painting classification using artistic comments," *Sensors*, vol. 21, no. 6, p. 1940, 2021.

[118] A. Efthymiou, S. Rudinac, M. Kackovic, M. Worring, and N. Wijnberg, "Graph neural networks for knowledge enhanced visual representation of paintings," *arXiv preprint arXiv:2105.08190*, 2021.

[119] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[120] B. Yang, X. Xiang, W. Kong, Y. Peng, and J. Yao, "Adaptive multi-task learning using lagrange multiplier for automatic art analysis," *Multimedia Tools and Applications*, vol. 81, no. 3, pp. 3715–3733, 2022.

[121] N. Garcia and G. Vogiatzis, "How to read paintings: Semantic art understanding with multi-modal retrieval," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[122] G. Castellano, V. Digeno, G. Sansaro, and G. Vessio, "Leveraging knowledge graphs and deep learning for automatic art analysis," *Knowledge-Based Systems*, vol. 248, p. 108 859, 2022.

[123] T. Standley, A. Zamir, D. Chen, L. Guibas, J. Malik, and S. Savarese, "Which tasks should be learned together in multi-task learning?" In *International Conference on Machine Learning*, PMLR, 2020, pp. 9120–9132.

[124] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *International conference on machine learning*, PMLR, 2018, pp. 794–803.

[125] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.

[126] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," *Advances in neural information processing systems*, vol. 31, 2018.

[127] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1871–1880.

[128] S. Chennupati, G. Sistu, S. Yogamani, and S. A Rawashdeh, "Multinet++: Multi-stream feature aggregation and geometric loss strategy for multi-task learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[129] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5824–5836, 2020.

[130] Z. Chen, J. Ngiam, Y. Huang, *et al.*, "Just pick a sign: Optimizing deep multitask models with gradient sign dropout," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2039–2050, 2020.

[131] L. Liu, Y. Li, Z. Kuang, *et al.*, "Towards impartial multi-task learning," ICLR, 2021.

[132] Z. Wang, Y. Tsvetkov, O. Firat, and Y. Cao, "Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models," *arXiv preprint arXiv:2010.05874*, 2020.

[133] B. Liu, X. Liu, X. Jin, P. Stone, and Q. Liu, "Conflict-averse gradient descent for multi-task learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 878–18 890, 2021.

[134] B. Lin, F. Ye, and Y. Zhang, "A closer look at loss weighting in multi-task learning," *arXiv preprint arXiv:2111.10603*, 2021.

[135] S. Gupta and J. Malik, "Visual semantic role labeling," *arXiv preprint arXiv:1505.04474*, 2015.

[136] P. Skalski, *Make Sense*, https://github.com/SkalskiP/make-sense/, 2019.

[137] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv*, 2018.