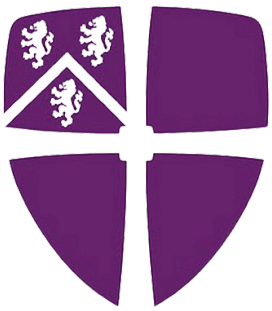
Attention Mechanism for Adaptive Feature Modelling

HAORAN DUAN

Attention Mechanism for Adaptive Feature Modelling



Haoran Duan

Department of Computer Science
Durham University

This dissertation is submitted for the degree of
Doctor of Philosophy

Ustinov College

February 2024

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university.

Haoran Duan
February 2024

Acknowledgements

Prior to anything else, I would like to express my profound gratitude to my supervisor, Dr. Yang Long, whose unmatched professionalism advanced my research immensely. I value not only his academic instruction but also his life lessons. I am incredibly appreciative of his invaluable support in regards to Equity, Diversity, and Inclusion (EDI), as he combined insightful scholarly advice with meaningful life teachings and compassionate support throughout my academic journey. He exemplified the uncommon combination of a mentor who enlightens both the intellect and the spirit, and is therefore deserving of high praise. I would also like to express my gratitude to Professor Ling Shao, Dr. Shidong Wang, Dr. Chris G. Willcocks, and Dr. Hubert P.H. Shum for their support, wisdom, and perseverance.

While pursuing a doctorate may appear to some as a solitary endeavor, I was fortunate to have many wonderful individuals accompany me on this voyage. To my cherished girlfriend, Zizhou Ouyang, your unspoken sacrifices, boundless support, and unfathomable love were the undercurrents that propelled my efforts to fruition. Your presence was a constant, gentle refuge from the turbulent storms of research and writing, making this voyage not only mine, but also ours. I would also like to thank my lab mates and collaborators for the laughter and joy throughout, as well as the professional discussions on each research project and endeavor - in particular, Dr. Zeyu Wang, Dr. Jie Su, Dr. Zhenyu Wen, Dr. Junyan Wang, Dr. Bingzhang Hu, Dr. Shuai Shao, Dr. Yang Bai, Dr. Bing Zhai, Chenghao Xiao, Fan Wan, Xingyu Miao, and Rui Sun. Additionally, I would like to thank the personnel at Durham University for their timely assistance throughout my studies.

Last but not least, my parents and family had to accept my separation from them while continuing to provide me with daily emotional and financial support. My affection and appreciation for them transcend the limitations of language. There are periods of quiet, determined, and arduous effort in every existence. I am reminded that a path of solitude precedes glory - times when I have fought, and endured for our aspirations. I am appreciative of my perseverance. I sincerely appreciate for the support and company of my family, friends, and collaborators listed above. If this journey is compared to traveling through a tunnel, and if I am the one penetrating the mountain, then they were unquestionably the lights that guided and accompanied me until I emerged into the light beyond.

Abstract

This thesis presents groundbreaking contributions in machine learning by exploring and advancing attention mechanisms within deep learning frameworks. We introduce innovative models and techniques that significantly enhance feature recognition and analysis in two key application areas: computer vision recognition and time series modeling. Our primary contributions include the development of a dual attention mechanism for crowd counting and the integration of supervised and unsupervised learning techniques for semi-supervised learning. Furthermore, we propose a novel Dynamic Unary Convolution in Transformer (DUCT) model for generalized visual recognition tasks, and investigate the efficacy of attention mechanisms in human activity recognition using time series data from wearable sensors based on the semi-supervised setting.

The capacity of humans to selectively focus on specific elements within complex scenes has long inspired machine learning research. Attention mechanisms, which dynamically modify weights to emphasize different input elements, are central to replicating this human perceptual ability in deep learning. These mechanisms have proven crucial in achieving significant advancements across various tasks.

In this thesis, we first provide a comprehensive review of the existing literature on attention mechanisms. We then introduce a dual attention mechanism for crowd counting, which employs both second-order and first-order attention to enhance spatial information processing and feature distinction. Additionally, we explore the convergence of supervised and unsupervised learning, focusing on a novel semi-supervised method that synergizes labeled and unlabeled data through an attention-driven recurrent unit and dual loss functions. This method aims to refine crowd counting in practical transportation scenarios.

Moreover, our research extends to a hybrid attention model for broader visual recognition challenges. By merging convolutional and transformer layers, this model adeptly handles multi-level features, where the DUCT modules play a pivotal role. We rigorously evaluate DUCT's performance across critical computer vision tasks. Finally, recognizing the significance of time series data in domains like health surveillance, we apply our proposed attention mechanism to human activity recognition, analyzing correlations between various daily activities to enhance the adaptability of deep learning frameworks to temporal dynamics.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Humans primarily perceive and interact with their surroundings through their sensory faculties. Our eyes constantly gather a plethora of information from the surroundings. Nevertheless, not all of this material is subjected to the same degree of meticulous analysis. The human perceptual system, a complex network consisting of the eyes and regions of the brain responsible for visual processing, use a method to effectively handle the vast amount of diverse information it receives [170]. This method is intricately linked to the concept of attention. The structure of our eyes is such that just a small region of the retina, called the fovea, contains a large number of photoreceptor cells[164]. The central zone captures fine details at the maximum level of resolution, while the surrounding periphery sections capture situations with progressively lower levels of resolution. When we desire to concentrate on particular aspects of a scene or item, we deliberately shift our gaze such that the object/part of interest aligns with our visual field[95].

Although attention mechanisms are primarily associated with perception, it is crucial to acknowledge that selective attention is a fundamental principle of information processing throughout the entire human brain[73]. Attention enables us to direct pertinent messages across various sensory channels in order to construct a unified comprehension of the environment. The process of selectively filtering information according to our behavioral objectives and requirements is crucial for effectively managing our finite cognitive capacities [38]. Neural attention processes actively control the flow of information across several sensory modalities. The attention mechanism enables us to focus our attention on important aspects of a target, while less significant elements are processed with reduced detail or disregarded altogether, by imitating the brain's ability to selectively attend to specific information [90].

Inspired by this inherent effectiveness of the human perception system, researchers have sought to incorporate similar strategies into deep artificial neural networks [239]. The idea is simple yet profound: if a model can learn to focus on the most pertinent parts of an

signals (e.g., images, time series) and ignore the irrelevant, it could potentially improve its performance while reducing computational cost. To this end, the attention mechanism can be conceptualized as a dynamic selection process [178]. Instead of processing all input features with equal importance, the model adaptively weights these features based on their relevance to the task at hand. This dynamic weighting allows the model to focus on crucial parts of the input while downplaying or ignoring less relevant ones. For instance, in an image classification task [227], while identifying a cat, an attention mechanism might give higher weight to features corresponding to the cat's whiskers, ears, and tail, and less weight to the background. Similarly, in time-series data modelling like human activity recognition, models prioritize critical time segments in wearable sensor data [168], like specific acceleration peaks during running, enhancing predictive accuracy by focusing on these informative data points amidst potential noise.

Developing and integrating attention mechanisms into deep learning models presents a unique set of challenges that can impact their efficacy and applicability across various tasks [68]. The primary hurdle lies in determining the optimal form and granularity of attention – discerning whether the focus should be on spatial regions, temporal segments, channels, or a combination thereof, and how fine or coarse this attention should be. To be more specific, one difficulty is determining what makes a region or feature "important" enough to warrant increased focus. Hard-coding attention based on human intuitions can fail to capture the true underlying statistics of the data. Implementing attention mechanisms with weak supervision presents significant challenges, often requiring extensive datasets. Researchers must carefully balance benefits against the additional computational burden of attention. Naively applying attention everywhere can be prohibitively expensive and sometimes unnecessary [41]. Efficient attention requires identifying which network components would benefit the most. There are also challenges in propagating attention-modulated signals through complex neural architectures. Attention learned for one task does not necessarily transfer well to other tasks. The field of deep learning has seen remarkable progress in developing attention mechanisms, yet key challenges persist. A primary challenge lies in designing attention models that balance accuracy with computational efficiency. For instance, non-local attention mechanisms, which excel in capturing long-range high-level semantic relations in data, face issues of scalability and high computational cost, especially when applied to high-resolution images or lengthy sequences. Furthermore, the generalizability of these mechanisms across different tasks and domains is not yet fully realized. Attention models that are effective in one application, such as video classification, may not necessarily yield similar results in another, like language translation. Finally, the development of flexible attention mechanisms capable of adapting to varied inputs and contexts is an ongoing research area. This includes

creating models that can dynamically alter their focus, for instance, by identifying and emphasizing salient features in complex datasets, a task that remains challenging in areas such as large-scale scene understanding or multimodal alignment.

The contributions of this thesis include overcoming the above challenges and proposing novel attention mechanisms for different tasks that achieve state-of-the-art performances. The rest of the thesis is organized as below:

- **Chapter 2 Background:** A comprehensive literature review will be provided from the deep learning paradigm and related models to the attention mechanism in this chapter. The introductions and related works for different applications are specifically included in each later chapter.

- **Chapter 3 Attention Enhanced Feature for Crowd Counting:** In this chapter, we investigate the attention mechanism in one of the crucial vision recognition tasks, which is crowd counting. We propose a dual attention mechanism (i.e., second-order and first-order attention) to enhance the representation learning of the multiscale crowd heads. The second-order statistics were extracted to retain the selectivity of the channel-wise spatial information for dense heads while first-order statistics, which can enhance the feature discrimination for the heads' areas, were used as complementary information. Via a multi-stream architecture, the proposed second/first-order statistics were learned and transformed into attention for robust representation refinement.

- **Chapter 4 Attention-based Feature Integration for Semi-Supervised Crowd Counting:** A huge part of the deep recognition models are trained based on the supervised paradigm, and only a small part of the attention-related research takes unlabelled data (e.g., unsupervised learning) into account. The unlabelled data is easy to collect, and it not only contains rich information for deep model training but also helps to reduce the cost of expensive human labor during data processing. Based on the investigation from Chapter 3, here in Chapter 4, we aim to propose an effective attention mechanism to take advantage of both labeled and unlabelled data for visual recognition. We also leverage the applications in Chapter 3, which we already built a good baseline, to evaluate our proposed attention mechanism. Recent works achieved promising crowd-counting performance but relied on the supervised paradigm with expensive crowd annotations. To alleviate the annotation cost in real-world transportation scenarios, in this work we proposed a semi-supervised learning framework, which can leverage both unlabeled/labeled data for robust crowd counting. We proposed an attention-driven crowd-driven recurrent unit along with two joint loss functions achieving better performance for crowd counting which uses the labeled data only.

- **Chapter 5 Unified Attention Model for Visual Feature Modelling:** After Chapter 4, we desire to consider how the attention mechanism can benefit general visual recognition tasks, hence we aim to propose a unified powerful attention-based deep model. We leveraged both the convolution operation and transformer layers (i.e., the most popular attention model) as the basis to design a hybrid attention model. Most importantly, we observed that the key rationale behind having a desired model is to handle the multi-level feature adaptively and dynamically. We propose two parallel modules along with multi-head self-attention to enhance the transformer. For local information, a dynamic local enhancement module leverages convolution to dynamically and explicitly enhance positive local patches and suppress the response to less informative ones. For mid-level structure, a novel unary co-occurrence excitation module utilizes convolution to actively search the local co-occurrence between patches. The parallel-designed Dynamic Unary Convolution in Transformer (DUCT) blocks are aggregated into a deep architecture, which is comprehensively evaluated across essential computer vision tasks.

- **Chapter 6 Attention Model for Dynamic Sequential Feature Modelling:** Visual data is indeed essential in various domains, but time series data holds a unique importance due to its temporal nature, making it indispensable for understanding dynamic processes and trends. Time series data captures sequential observations over time, making it invaluable for tasks like stock price prediction, climate forecasting, and health monitoring. Its temporal dependencies and seasonality patterns necessitate specialized modelling techniques like RNNs and LSTM, which excel at capturing these nuances. Time series data is also frequently encountered in decision-making scenarios, where historical trends and future predictions play a critical role. Therefore, while visual data offers spatial insights, time series data provides the temporal dimension necessary for making informed and timely decisions, making it an essential modality in machine learning and deep learning. In this chapter, we investigate a novel attention mechanism for the human activity recognition tasks, where the time series data is collected from the wearable sensors. Specifically, the proposed attention mechanism is utilised to adaptively learn the feature correlations for different activities alleviating the activity intrusion problems.

Publication Lists

The works, which both mainly contributed to this thesis and loosely related to the topic of this thesis, have been previously published in the following peer-reviewed publications by the author, and listed below. Incorporating a comprehensive list of my publications underscores

my academic journey, offering a transparent, verifiable, and comprehensive record of my scholarly contributions and expertise in the field throughout my PhD study.

The two principal directions of these publications, computer vision modelling and temporal information integration, are significantly enhanced by the application of attention mechanisms. Within computer vision modeling, attention mechanisms are instrumental in advancing tasks such as crowd counting, where they enable precise quantification of individuals in diverse environments, overcoming challenges like varying densities and obstructions. They also play a crucial role in general vision tasks, such as image segmentation, by focusing on relevant image regions to improve boundary delineation accuracy, and in image enhancement, by identifying and amplifying or correcting key image features, thereby elevating image quality. By doing these works, I obtained a more profound understanding of the knowledge of attention mechanism design and digital visual processing. Moreover, these works specifically helped me form the dissertation within Chapter 3, Chapter 4 and Chapter 5.

- Duan, H., Long, Y., Wang S., Zhang H., Willcocks C., Shao L., 2023 **Dynamic Unary Convolution in Transformers.** IEEE Transactions on Pattern Analysis and Machine Intelligence.

- Duan, H., Wang S., Guan Y., **SOFA-Net: Second-Order and First-order Attention Network for Crowd Counting.** British Machine Vision Conference (BMVC 2020).

- Wang, Z., Li, X., Zhao, L., Duan, H., Wang, S., Liu, H. and Zhang, X., 2023. **When Multi-Focus Image Fusion Networks Meet Traditional Edge-Preservation Technology.** International Journal of Computer Vision, pp.1-24.

- Gao, R., Wan, F., Organisciak, D., Pu, J., Duan, H., Zhang, P., Hou, X. and Long, Y., 2023, July. **Privacy-Enhanced Zero-Shot Learning via Data-Free Knowledge Transfer.** In 2023 IEEE International Conference on Multimedia and Expo (ICME) (pp. 432-437). IEEE.

- Wang, Z., Li, X., Duan, H. and Zhang, X., 2022. **A self-supervised residual feature learning model for multifocus image fusion.** IEEE Transactions on Image Processing, 31, pp.4527-4542.

- Duan, H., Shum, H.P., Hu, B.,Guan, Y., Wang, S., Long, Y., 2023. **Taking advantage of unlabelled data for Semi-Supervised Crowd Counting.**

In the domain of temporal information integration, the impact of attention mechanisms is equally profound. In video analysis, attention mechanisms focus on pertinent frames or

segments over time, facilitating a deeper understanding of dynamic scenes and interactions. This capability is crucial for integrating essential temporal features, ensuring a comprehensive understanding of context and event progression in videos. For human activity recognition, these mechanisms are pivotal in identifying subtle temporal patterns and movements, leading to a more nuanced and accurate recognition and cost saving. These works not only helped me to understand how to apply the attention mechanism for integrating temporal information but also helped to form the Chapter 6 in my dissertation.

- Miao, X., Bai, Y., Duan, H., Huang, Y., Wan, F., Xu, X., Long, Y. and Zheng, Y., 2023. **DS-Depth: Dynamic and Static Depth Estimation via a Fusion Cost Volume.** IEEE Transactions on Circuits and Systems for Video Technology.

- Wan, F., Wang, J., Duan, H., Song, Y., Pagnucco, M. and Long, Y., 2023, June. **Community-Aware Federated Video Summarization.** In 2023 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

- Wang, R., Wei, Z., Duan, H., Ji, S., Long, Y. and Hong, Z., 2022. **EfficientTDNN: efficient architecture search for speaker recognition.** IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, pp.2267-2279.

- Duan, H., Wang, S., Wang, SZ., Huang, Y., Long, Y., Zheng, Y., 2023. **Wearable-based Behaviour Interpolation for Semi-supervised Human Activity Recognition.**

Chapter 2

Background

2.1 Deep Learning

Deep learning has become one of the most popular terms in the past decade, a part of machine learning methods. Media computing tasks based on deep learning have not only attracted widespread attention in academia but have also been equally in-depth in development and application in the industrial field [59]. Among them, low-level intelligent media processing technologies based on learning, such as image and video compression, denoising, and enhancement, as well as high-level intelligent analysis technologies, such as face recognition, pedestrian re-identification [187], target detection [141] and tracking [32], action recognition, etc. It has deeply penetrated into our lives and production and provided us with various conveniences. However, despite these achievements, an obvious gap persists in information processing and comprehension abilities between machines and humans. The key challenge in contemporary artificial intelligence multimedia technology development revolves around bridging this gap, with the primary objective being the enhancement of machine information processing efficiency and adaptability to a level commensurate with human capabilities [239].

2.1.1 Learning Paradigm

Supervised learning is a widely employed methodology in the fields of deep learning and machine learning. Supervised learning is a method of teaching a model by giving it a large amount of data that is labeled, meaning each input data point is paired with its associated output label or target [17]. In regression questions, the target can be represented as a numerical value, while in classification problems, it is represented as a class label. The primary objective of the model is to acquire or establish the mapping correlation between the given input data and the output labels, which can be considered as a function. This enables the

model to reliably anticipate the associated output when fresh, unknown data is provided to it. The subsequent steps outline the usual procedure for doing supervised learning [239]. Data Collection and Processing: Acquire data for the purpose of training and validating the model, often encompassing input attributes and goal output. Data typically requires preprocessing, including normalization, denoising, and encoding. Model Selection: Select a suitable model (such as decision tree [104], neural network [157, 74], or support vector machine [76]) based on the intricacy of the problem and the attributes of the data. Model Training: The process of training the model involves using labeled training data, which consists of input data and corresponding target output. This stage typically entails fine-tuning the model's parameters to optimize its ability to learn and accurately translate inputs to outputs. Typically, this is accomplished by minimizing a loss function that quantifies the difference between the predictions made by the model and the actual targets. Model Validation and Testing: Assess the model's performance on separate validation and/or test data sets to verify its capacity to generalize, meaning its ability to make accurate predictions on new, unknown data. Model Deployment and Application: After the model has been confirmed as successful, it can be implemented in a real-world setting to provide predictions on fresh data.

Unsupervised learning in the context of deep learning refers to a technique used to train a model to identify hidden structures and patterns in input data, without depending on external labels or classification information [112]. Unsupervised learning involves enabling the model to autonomously understand the distribution and inherent relationships within the data, without relying on labels to guide or adjust its learning process. During this process of learning, the network autonomously organizes itself and progressively acquires the ability to represent data in a more efficient manner, namely through learning an effective data representation. During a typical unsupervised learning activity, such as clustering [238], the dataset we are given lacks labels, indicating that we are unaware of the categories to which each data point corresponds. The algorithm must autonomously identify the structure or pattern of the data using an inherent similarity or distance metric. This involves classifying the data in a way that minimizes the differences between data points within the same category and maximizes the differences between data points in different categories. This approach is extensively employed in several fields such as market segmentation [225], social network analysis[148], computational biology [156], image segmentation [147], and others. It is particularly useful when the actual labels or categories of data are unknown, and the algorithm must establish logical classifications or groups based on data distribution or similarity. The fundamental purpose of unsupervised learning is not to instruct the computer on how to perform tasks, but rather to enable the computer to autonomously acquire knowledge and uncover concealed principles, frameworks, and patterns from extensive datasets. This learning

approach is highly advantageous in several situations, particularly when there is no distinct anticipated outcome or when labeled data is limited and costly. Unsupervised learning encompasses more than just clustering. It encompasses a range of problems including density estimation [135], dimensionality reduction [176], and generative model development[60]. Unsupervised learning algorithms must navigate the realm of data without explicit direction, endeavor to comprehend the underlying processes that generate the data, and apprehend the fundamental characteristics and structures of the data.

Semi-supervised learning is a machine learning approach that falls between supervised learning and unsupervised learning [209]. Semi-supervised learning involves training a model using a combination of partially labeled and partially unlabeled training data. This method specifically utilizes a limited quantity of labeled data and a substantial quantity of unlabeled data, simultaneously acquiring explicit patterns from the labeled data and latent structures from the unlabeled data during the training process. Acquiring unlabeled data is often convenient and cost-effective in various real-world scenarios, but acquiring labeled data might be costly or impractical due to the requirement for specialized expertise or human involvement. Semi-supervised learning arises in this environment with the goal of constructing a more robust and comprehensive learning model by integrating both labeled and unlabeled input. Semi-supervised learning generally includes the following main techniques. Self-training [167]: The model is first trained using labeled data. It then utilizes this trained model to make predictions on unlabeled data. The predictions are then used as new labels. The data that the model predicts with high confidence is added to the training set. This process is repeated iteratively for further training. Multi-view Learning [92] refers to the process of leveraging data that can be viewed from numerous perspectives or feature spaces. The objective of multi-view learning approaches is to improve learning performance by effectively integrating information from these diverse viewpoints. Generative Models [100]: This strategy aims to understand the underlying mechanism that generates the data by determining the joint distribution of labeled and unlabeled data. This knowledge is then utilized to improve the accuracy of predictions. Graph-based semi-supervised learning [101] involves the transformation of data into a graph format, where nodes represent individual data points and edges represent the similarity between these points. The objective of the model is to utilize the graph structure to propagate and refine label information, ultimately predicting the labels of unlabeled data. Semi-supervised learning is extensively utilized in many fields, including image identification, natural language processing, and bioinformatics, among others. By effectively utilizing a small set of labeled data with a substantial quantity of unlabeled data, it is possible to enhance learning performance and accuracy to the fullest extent. Semi-supervised learning is frequently employed in actual problem-solving scenarios

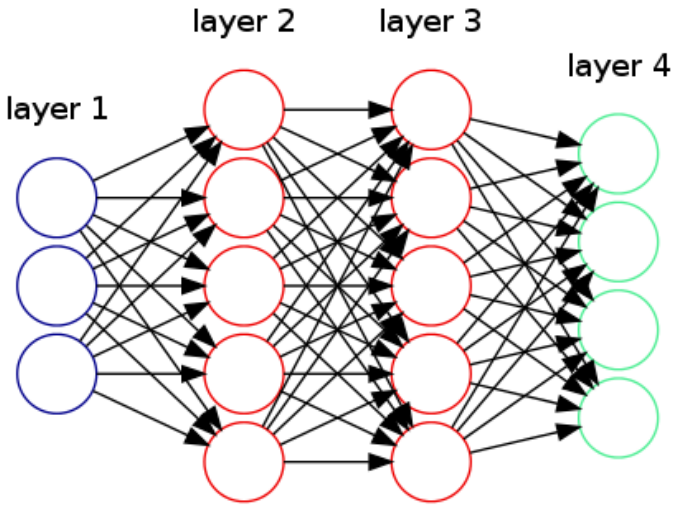


Fig. 2.1 Example visualization of MLP[123].

as it offers a comparatively efficient and viable solution, particularly when there is a shortage of labeled data.

2.1.2 Deep Models

When exploring models in the field of deep learning, the Multilayer Perceptron (MLP) is commonly used as a foundational starting point. MLP, or Multi-Layer Perceptron, is a type of neural network that uses many layers of neurons to perform nonlinear transformations on information. Its goal is to learn and approximate complex mapping relationships. A Multilayer Perceptron (MLP) [119] consists of several layers: the input layer receives a feature vector as input, followed by a sequence of hidden layers that process and encapsulate the input information, and ultimately, the output layer generates the predictive outcomes of the model. Neurons in the model are coupled through weights within each layer, and activation functions (such as ReLU, Sigmoid, etc.) allow the model to handle nonlinearities. The model utilizes the backpropagation algorithm and gradient descent approach to iteratively update the network weights based on prediction mistakes, allowing it to understand the underlying patterns in the data.

Expanding on the groundwork established by MLP, Convolutional Neural Networks (CNN) [123] introduced groundbreaking advancements to the domain of deep learning, particularly in the areas of image processing and computer vision. CNNs effectively capture local spatial structure information in picture data by utilizing processes such as local receptive fields, weight sharing, and spatial pooling. The essential elements of this model consist of

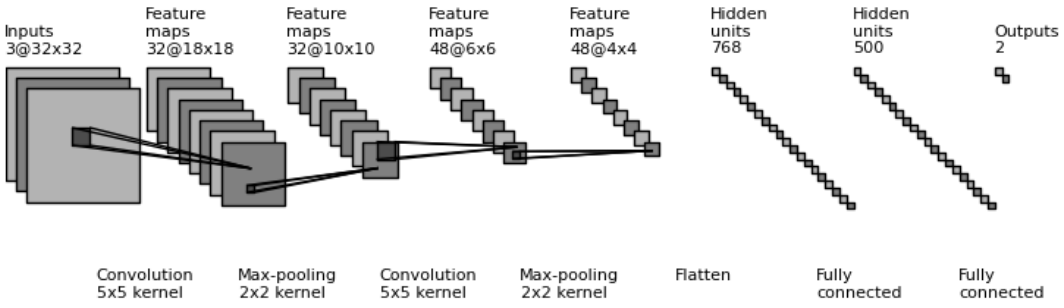


Fig. 2.2 Example visualization of Convolution Neural Network[106].

convolutional layers and pooling layers. The convolutional layer performs a scan of the input feature map using filters (or kernels), resulting in localized feature extraction and weight sharing, which effectively reduces the parameter size of the model. The pooling layer reduces the complexity of the spatial arrangement in the feature map by performing downsampling operations. Its purpose is to extract significant features and reduce computational load. The two layers are stacked in an alternating manner to create a complex structure that can effectively represent image content at different levels of abstraction. This design achieves exceptional performance in numerous computer vision applications.

Here, we highlight three distinct and popular network architectures built based on the blocks mentioned above, which are ResNets, Capsule Network and Generative Adversarial Network. Residual Networks, or ResNets [74], represent a significant advancement in convolutional neural network design, specifically intended to facilitate the training of extremely deep networks. Traditional deep networks encounter difficulties such as vanishing and exploding gradients as they grow deeper, but ResNets address this with an innovative architecture incorporating "skip connections" or "shortcut connections." These connections allow gradients to bypass one or more layers, enabling identity mapping where the outputs of the skipped layers are added to the layers' outputs. The introduction of residual blocks, the core components where these shortcut connections are implemented, has been pivotal. ResNets have been constructed with over 100 layers, far deeper than previous models, and have demonstrated remarkable performance improvements in various tasks, especially image recognition and classification. Capsule Networks, conceptualized by Geoffrey Hinton and his collaborators [180], introduce an innovative approach to neural network architecture, aiming to enhance the processing of spatial hierarchies and relationships within data, an area where traditional convolutional neural networks (CNNs) typically fall short. The key innovation in Capsule Networks is the introduction of "capsules" clusters of neurons that represent various properties of the same entity inside convolution, such as its position, size, and orientation.

These capsules employ a dynamic routing mechanism, allowing for the communication and agreement on the presence of higher-level entities. This architecture is specifically adept at preserving detailed spatial relationships, which significantly enhances its ability to recognize objects across diverse viewpoints and poses. Capsule Networks are particularly notable for their potential in tasks requiring an understanding of complex spatial relationships and have shown promise. Generative Adversarial Networks (GANs) [60], introduced by Ian Goodfellow and his team, are a groundbreaking class of artificial intelligence algorithms within the realm of unsupervised machine learning. Comprising two convolution neural networks, the generator and the discriminator, which operate in a competitive manner, GANs have revolutionized several aspects of machine learning and data generation. The generator creates new data instances that mimic a real data distribution, while the discriminator evaluates the authenticity of these instances, discerning whether they are from the actual dataset or fabricated by the generator. This adversarial training process, where the generator aims to maximize the error rate of the discriminator as the discriminator seeks to minimize its own error rate, has proven highly effective. GANs have made a substantial impact in fields such as image generation and style transfer. They are particularly noted for their ability to produce highly realistic images and have found applications in art, photo editing, and the generation of synthetic datasets for further model training.

When it comes to handling temporal data, Recurrent Neural Networks (RNN) [185] are the preferred model due to its internal looping connections. The fundamental characteristic of RNN is its capacity to preserve and exploit concealed state information from preceding time points in order to impact the output at the subsequent time point. The ability of RNN to comprehend and use the dynamic attributes included in temporal data is facilitated by this sequence-dependent feature. While RNNs have the potential to capture dependencies in lengthy sequences, they encounter significant obstacles in effectively learning long-term dependencies due to the presence of vanishing and ballooning gradient issues.

Each of the three network models mentioned above has distinct characteristics and applications, which serve as the basis for the advancement and utilization of deep learning technology. Furthermore, they enhance our comprehension and ability to analyze intricate data patterns across several dimensions. These models have shown outstanding performance in their specific areas and have become essential technologies in numerous research and application initiatives.

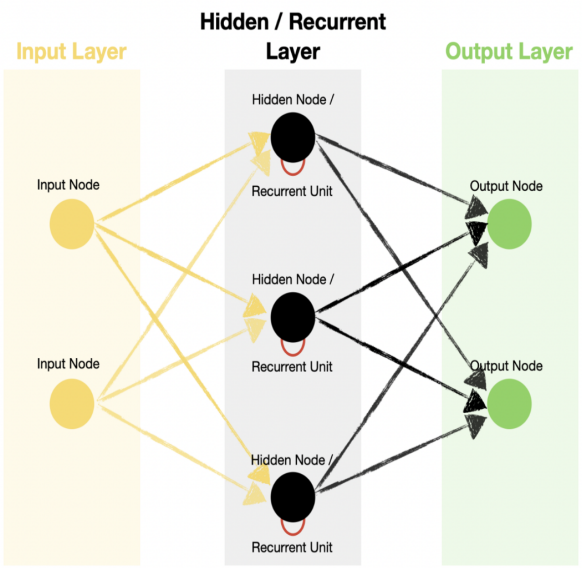


Fig. 2.3 Example visualization of Recurrent Neural Networks[185].

2.2 Attention Mechanisms

Following the introduction of foundational deep learning models that provided effective tools for machine learning, researchers shifted their focus towards tackling more intricate and nuanced challenges. These include managing diverse data with long-term relationships, hierarchical structures, and varying levels of significance. The introduction of the Attention Mechanism into models was a response to these issues [72]. The fundamental concept behind the attention mechanism is to enhance the model's capability to choose and concentrate on crucial segments during the processing of incoming data, such as sequences or images. This involves assigning greater importance, or weights, to certain segments. In tasks such as machine translation or text summarization, the contribution of distinct input parts to the output varies when working with sequential data. The attention mechanism enables the model to learn dynamic weight allocations for important words or phrases, enhancing its precision and flexibility in resolving tasks that involve understanding complex relationships in the input data [98]. Integrating the attention mechanism allows models to prioritize important information in the inputs, leading to improved performance and interpretability in tasks such as natural language processing and computer vision.

The concept of an attention mechanism encompasses the ability to selectively focus on and process information with varying degrees of importance, whether in the context of human cognition or machine learning. Attention mechanisms exhibit a variety of types, roles, and applications. Drawing inspiration from studies in neuroscience and psychology,

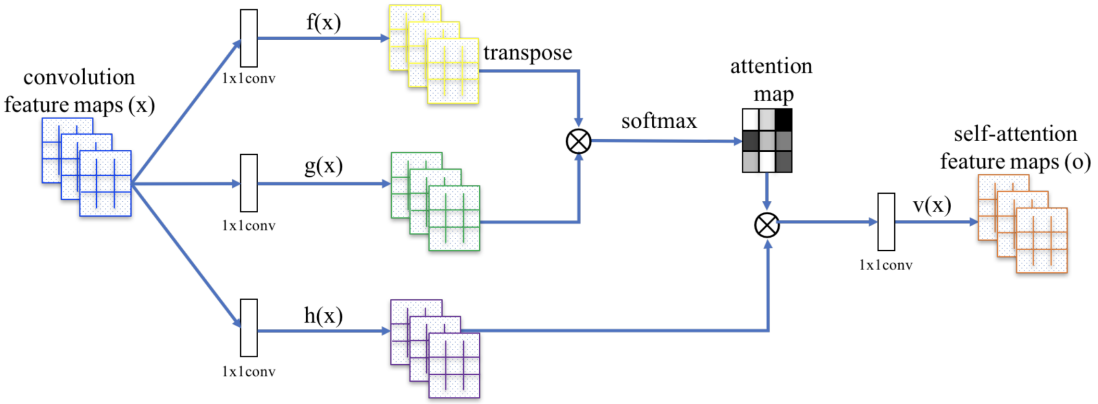


Fig. 2.4 Example visualization of Self-Attention Model [256]. The f, g, h, v denotes the intermedia calculations and the x is the input.

these mechanisms aim to replicate the way humans allocate their attention in perceptual and cognitive processes. In essence, attention mechanisms enable the human perceptual system to concentrate on localized stimulus feedback, making them an integral component of the cognitive process. Human attention enhances the sensitivity of input signals at the level of cellular synapses, thereby refining the precision of these signals. It achieves this by selectively enhancing the transmission of vital signals while concurrently reducing noise levels, ultimately reshaping neural sensations [210].

In the field of attention mechanisms, research has a long and storied history, with its core ideas tracing back to the 1990s. In 2014, Google's DeepMind team successfully combined attention mechanisms with Recurrent Neural Networks (RNNs) and applied them to image classification tasks, achieving remarkable performance and sparking widespread interest in attention mechanisms [151]. RNNs are particularly effective in tasks where context is important, as they can process inputs in a sequence, one at a time, and retain information from previous inputs. This is achieved through the use of hidden layers that retain a state or memory of the past inputs. However, one challenge with RNNs is the vanishing gradient problem, where gradients can become extremely small during backpropagation, making it difficult to learn and adjust the weights of the earlier layers in the network. Subsequently, Bahdanau and others introduced attention mechanisms into the field of Natural Language Processing (NLP), enabling joint learning of translation and alignment and expanding the scope of attention mechanism applications [4]. Soon after, Xu and his colleagues applied attention mechanisms to image captioning tasks, demonstrating their effectiveness across different data types [237]. In 2017, Google's research team introduced a novel self-attention

mechanism in machine translation, further diversifying the landscape of attention mechanisms and injecting fresh vitality into related research [210].

2.2.1 Selective Attention Mechanism

The selection attention mechanism is an explicit attention method that assesses the significance of various data components for task optimization. It calculates attention weights and uses them to selectively enhance important data features while suppressing irrelevant ones. This process involves computing a set of weights, typically through a softmax function, which are then applied to the input features to amplify those that are more relevant to the task at hand. The underlying idea is to mimic cognitive attention in humans, where focus is directed towards certain aspects of the input while ignoring others. Selective attention processes are extensively utilized in diverse information processing and analysis activities, including language signals and visual signals. This significantly enhances the nonlinear expressive capacity of neural networks and their capability to extract high-level semantics. By focusing on specific parts of the data, these networks can more effectively learn complex patterns and relationships, leading to more accurate predictions or classifications.

The selective attention mechanism utilizes neural networks to process the data or features and build an attention mask. This attention mask predicts the significance of different parts of the data or features, which is then used to amplify or diminish specific elements. The attention mask is typically generated through a learnable neural network layer, which assesses each part of the input data and assigns a relevance score. These scores are then normalized and used to scale the input features, thereby altering the network's focus.

Selective attention techniques have diverse uses in various network models such as CNNs, RNNs, and GNNs, allowing for adaptability in handling varied problems. For example, in Convolutional Neural Networks (CNNs), CBAM [227] and SE [82] attention mechanisms serve as modular components that can be easily integrated into the network. CBAM sequentially applies two distinct attention modules: the channel attention module, which focuses on 'what' is meaningful by emphasizing certain channels of the feature map, and the spatial attention module, which concentrates on 'where' is an important aspect in the feature map. This dual focus allows CBAM to refine the feature maps adaptively, making it highly effective in tasks like image classification and object detection. The SE mechanism adopts a different approach. It globally aggregates the spatial information of the feature maps into channel-wise statistics. This process involves squeezing the global spatial information into a channel descriptor, which is then used to recalibrate the feature channels by explicitly modeling the interdependencies between them. The recalibrated channels are then re-weighted to enhance the representational power of the network. These

mechanisms enhance the representation of features in both spatial and channel dimensions, hence improving performance across many tasks. For instance, in image recognition, these attention mechanisms help the network to focus more on the parts of the images that are more relevant to the identification task, such as focusing on the features of a bird in a complex background for species classification. In object detection, they assist in isolating and highlighting objects of interest amidst a cluttered scene. In addition, their attention outputs can be utilized as inputs to aid in comprehending complex crowd behavior or improving visual representations for reinforcement learning models [153], enhancing the agent's ability to make decisions based on a more accurate and focused analysis of the visual input. For example, in reinforcement learning scenarios involving navigation or interaction within an environment, these mechanisms can help in identifying and focusing on key elements in the scene, like obstacles or targets, thus enabling more strategic and efficient decision-making.

Convolutional LSTM [3] is a type of Recurrent Neural Network (RNN) that uses convolutional layers within LSTM modules to learn filter coefficients. This approach allows for the direct and extensive application of selective attention processes. The integration of attention modules with LSTM modules enhances the representation of motion information and improves the accuracy of action recognition tests. In addition, attention modules have the ability to dynamically adjust input signals in recurrent neural networks (RNNs), allowing for selective focus.

These examples demonstrate how selective attention techniques can improve the adaptability and performance of neural network models in various tasks.

2.2.2 Self-Attention Mechanism

The self-attention mechanism [210], a pivotal component in the Transformer model, revolutionizes information processing by aligning internal and exterior information, thereby refining the precision of local feature representations. This mechanism operates on the principle of 'non-local averaging,' [229] which allows the network to assess and integrate information across all parts of the input data. At its core, self-attention involves three key elements: Queries, Keys, and Values. Each element in the input data is transformed into these three representations. The Query corresponds to the specific element being focused on, while the Keys represent all elements in the input data, including the element corresponding to the Query. The process begins by computing the similarity between the Query and each Key, typically using a scaled dot-product. This similarity measure reflects the relevance of each element in the context of the Query. Following the similarity computation, a softmax function is applied to these scores, converting them into a probability distribution. This distribution effectively weighs the importance of each element in the input when considering

the one represented by the Query. Subsequently, these weights are applied to the Values, which carry the actual content of the input data. The weighted sum of these Values, based on the computed attention scores, generates a new representation for each input element. This new representation is a blend of information from the entire input sequence, weighted by their relevance to the specific element in question.The result of this process is a series of output elements where each element is informed by a contextually weighted combination of all other elements in the input sequence. This allows the model to capture intricate interdependencies and relationships within the data, regardless of their position in the sequence. In essence, self-attention mechanisms grant neural networks a more dynamic and contextual capability to process and represent information, significantly enhancing tasks that involve complex dependencies.

The self-attention mechanism was initially implemented in the field of natural language processing with the purpose of acquiring semantic characteristics that correlate to various combinations of word embeddings. Subsequently, it was utilized in a wider range of image and video processing endeavors, including pedestrian identification, tracking multiple objects, and improving image quality. The objective of this is to utilize information from beyond the immediate area to improve the portrayal of features within specific local regions. Currently, self-attention is being used in various jobs.

Graph Attention Networks [211] have incorporated the self-attention mechanism into graph models to regulate the connections between nodes in the graph, hence improving the precision of graph classification tasks. VL-Bert [199] is a flexible model that utilizes self-attention and is used for several tasks that include the combination of visual and language information. VL-Bert operates by concurrently processing visual inputs, such as images, and textual inputs through a unified architecture based on the Transformer model. It transforms visual inputs into a format compatible with textual tokens by extracting features using a pre-trained CNN. These features are then assimilated with textual data, enabling the model to simultaneously focus on and interpret elements from both visual and textual inputs. This process is crucial for tasks that demand an understanding of the complex interplay between visual and linguistic elements, such as visual commonsense reasoning [249], visual question-answering [2], and image captioning [79]. In these applications, VL-Bert's extended self-attention mechanism comes to the fore, understanding not just the relationships within textual data, but also between text and visual elements. For instance, in visual question-answering tasks, the model integrates visual cues from an image with a textual query to generate accurate responses. In image captioning, it crafts descriptive, contextually relevant text based on the visual information. This dual-data processing ability, fortified by extensive pre-training on large visual-language datasets, significantly enhances VL-Bert's performance

in these tasks, showcasing its superiority in understanding and integrating visual and textual data. A recent study has thoroughly investigated the utilization and function of self-attention mechanisms in activities related to comprehending videos.

Furthermore, some studies have investigated the process and structure of self-attention to enhance its efficacy [210]. Scientists have categorized attention weights in self-attention into four distinct information composition patterns: (1) the combination of Query and Key content, (2) the combination of Query content and the relative position of Query and Key, (3) the content of Key alone, and (4) the relative position of Query and Key. Studies on spatial self-attention processes have conducted empirical research to examine the impact of various forms of information included in these patterns. Interestingly, in conventional self-attention models, it was shown that attention weights were not dependent on the selection of target placements. Given this discovery, certain researchers have undertaken the task of reconfiguring self-attention mechanisms, suggesting more efficient options to decrease computational complexity. In addition, they have separated self-attention into a unary word that signifies the current location information and a binary term that signifies the link between two positions, hence decreasing the complexity of optimization.

Recently, researchers have integrated these two distinct forms of attention systems. Through the fusion of these systems, self-attention modules effectively utilize non-local information in order to learn selective attention masks. These strategies improve the module's receptive field, enabling it to make more knowledgeable judgments when predicting the significance of information. Nevertheless, these techniques have a tendency to augment the number of parameters and computational intricacy of attention models, due to the incorporation of fully linked networks and self-attention modules. Consequently, optimizing neural networks becomes more arduous. In addition, these attention approaches fail to take into account the intrinsic links between data or features when assessing their significance.

Chapter 3

Attention Enhanced Feature for Crowd Counting

In this chapter, we aim to investigate how the attention mechanism helps the vision task crowd counting. Automated crowd counting from images/videos has attracted more attention in recent years because of its wide application in smart cities. But modelling the dense crowd heads is challenging and most of the existing works become less reliable. To obtain the appropriate crowd representation, in this work we proposed SOFA-Net(Second-Order and First-order Attention Network): second-order statistics were extracted to retain selectivity of the channel-wise spatial information for dense heads while first-order statistics, which can enhance the feature discrimination for the heads' areas, were used as complementary information. Via a multi-stream architecture, the proposed second/first-order statistics were learned and transformed into attention for robust representation refinement. We evaluated our method on four public datasets and the performance reached state-of-the-art on most of them. Extensive experiments were also conducted to study the components in the proposed SOFA-Net, and the results suggested the high-capability of second/first-order statistics on modelling crowd in challenging scenarios. To the best of our knowledge, we are the first work to explore the second/first-order statistics for crowd counting.

3.1 Introduction

Crowd counting aims to count the number of people in images or videos of crowd scenes. It plays a pivotal role in real-world applications such as video surveillance, traffic planning, public security, etc. Earlier attempts were based on pedestrian detection [213] or human segmentation [262] in crowd. Recently, crowd counting has been regarded as an image-based

density map regression task, and counting can then be conducted through integrating the densities. Density-map based methods achieved promising counting results in crowded scenes when it's difficult to detect subjects due to distance, occlusions, etc.

Previous density map regression works [87, 166] were proposed to learn the regional objects mapping, and recently Convolution Neural Network (CNN) became the major technique for crowd representation learning. In [183], a switching CNN learning inherent structural and functional differences is proposed to tackle large scale and perspective variations in crowd counting. Due to the diverse number of subjects and the various dense or sparse crowd patterns, most recent CNN-based approaches [16, 121, 125, 128, 236] were proposed to estimate density maps by handling the multi-scale problems in crowd scenes. However, these methods become less reliable when the areas of pedestrians' heads are dense and very small. In [33], it was found that high density areas tended to be underestimated, while the low density areas tended to be overestimated. This observation suggested that a better representation should be learned in such challenging crowd scenarios. So we proposed a deep Second-Order and First-order Attention Network (SOFA-Net) for crowd modelling. Second-order statistics learning was successfully used to improve the representation learning [117, 42, 31, 230] or to recognize small objects in remote sensing [31, 221]. In this work, the second-order statistics leads our model to learn robust crowd representation by retaining selectivity of spatial information. First-order statistics, which can capture the discriminated spatial characteristic for crowd, was also used as complementary information. A Statistic-Wise Convolution operation was also proposed to effectively transform the second/first-order statistics into attentions for our network. Then a deep attention architecture was designed to handle multiple feature streams for generating the crowd density maps. For better generation quality [269, 125], a normalization strategy and a scale enhancement were also used. Our main contribution can be summarized as:

- To the best of our knowledge, this is the first work proposed to use second/first-order statistics for crowd modelling. We analyse the effects of second/first-order statistics for crowd counting qualitatively and quantitatively. Then, the overall experimental results suggested their feasibility in challenging crowd scenarios.
- We designed a multi-stream architecture with a Statistic-Wise Convolution to learn the second/first-order statistical attentions for crowd density map generation. Also, several tailored components were also proposed and evaluated.
- We tested our method on four popular public datasets, and it reached the state-of-the-art performance on most challenging datasets.

3.2 Related Work

Crowd Counting Early methods were based on designing detection/segmentation algorithms [213, 262], yet these methods may be heavily affected by occlusions, making them less practical. In [115], density map estimation approach was first introduced, which aimed to identify the centre locations of the subjects to avoid the error-prone detection procedures. CNN-based methods were the main techniques for representation learning in crowds counting [130, 125, 193, 252]. A maximum-excess-over-pixel loss was proposed with regional feature pattern to utilize the spatial information to count people in different density levels [33]. In [128], Liu et al. proposed a structured feature enhancement module by conditional random fields with a dilated multi-scale structural similarity loss to adapt the scale variations. Zhang et al. present MCNN [258], featuring three separate convolutional network columns. Each column is distinct in its convolution kernel count and dimensions. The training process involves independently training each column, followed by a combined fine-tuning. The model then uses 1X1 convolution kernels to merge features from the three scales, creating a density map. This architecture is flexibility with input image sizes is an advantage, as it omits fully connected layers. Notably, the team also introduced the ShanghaiTech dataset, a significant contribution to crowd counting research with its extensive data and benchmarking. Dong et al. propose MMNet [48], a scale-aware, end-to-end network. Their approach not only utilizes multi-scale features from varied filter sizes but also combines features from different stages to adapt effectively to the varying scales of human heads for accurate crowd counting. Gao et al. [55] observe that previous work mainly targets local visual details of crowds and often neglect essential contextual and attention data. To overcome this, they developed the SCAR framework (Spatial-Channel-wise Attention Regression Network), which consists of two innovative models: SAM (Spatial-wise Attention Model) and CAM (Channel-wise Attention Model). SAM's role is to encode the full input image, gathering extensive context information for enhancing the accuracy of the density map predictions. Meanwhile, CAM zeroes in on the most impactful channel features, thereby bolstering the network's ability to handle noisy and complex backgrounds. The fusion of data from these two attention-focused models results in a more refined and accurate density map. Although these methods achieved good performance, it is still uncertain how to statistically select the learnt visual features for crowd counting, which is presented in this chapter.

It is also worth to mentioning that face/head detection [46] and crowd counting are very similar tasks yet still be distinct among each other with unique challenges. Face/head detection aims to identify and precisely locate individual faces or heads, often dealing with varied scales and orientations, and requiring detailed recognition against diverse backgrounds. In contrast, crowd counting focuses on estimating the number of people in highly dense

and cluttered scenes, where individual detection is less critical, and exact localization gives way to aggregate count estimation [219]. The major challenges in crowd counting include handling extreme densities with severe occlusions, managing perspective distortions where the scale of individuals varies significantly, ensuring robustness across diverse environmental conditions, achieving real-time processing capabilities for applications like surveillance, and maintaining privacy by focusing on counts rather than individual identification. These challenges necessitate advanced deep learning approaches, such as density estimation models, perspective-aware algorithms, extensive data augmentation, model optimization for speed, and privacy-preserving data practices, differentiating crowd counting significantly from the more localization-focused task of face/head detection.

Second/First-Order Statistic In large scale CNN network, the global average or max pooling was normally set at the end (as the first-order pooling) to capture the image representation by first-order statistical summary[192]. Recently, second-order statistics were also explored for improving the representation learning ability in many computer vision tasks [117, 42, 31, 230]. Li et.al evaluated the effectiveness of second-order information for large scale visual recognition with a trainable matrix power normalized covariance pooling [117]. The combination of first-order and second-order information was also employed in a multi-level architecture CNN [42] to improve the image texture discrimination. Based on second-order statistics, a recurrent transformer network was proposed [31] to learn transformation-invariant representation for remote sensing with great performance at recognizing small objects.

3.3 Methodology

3.3.1 Problem Statement

Following [115, 87, 259, 88, 33], we describe the crowd counting problem as follows:

Given a crowd image $\mathbf{X}$ with c pedestrians' heads $\mathbf{H} = \{\mathbf{h}_i \in \mathbb{W}^2\}_{i=1}^{c}$, where $\mathbf{h}_i$ is the x-y coordinate of the i_{th} center point of the subject's head and $\mathbb{W}^2$ denotes the two dimensional positive real values. The (ground truth) density map can be constructed by c Gaussian function $\mathscr{N}$ over all the heads' pixel grids in image $\mathbf{X}$, such that the crowd counts can be calculated by the integrals of density map. The (ground truth) density map can be written as:

$$\mathbf{D}^{gt} = \sum_{\mathbf{h}\in\mathbf{H}} \mathscr{N}(p; \mu = \mathbf{h}, \sigma^2) \tag{3.1}$$

where $p \in \mathbf{X}$ denotes the image pixels, gt is the superscript of the ground truth, pr is the superscript of the prediction and σ is a very small number (spanning a few pixels [115]).

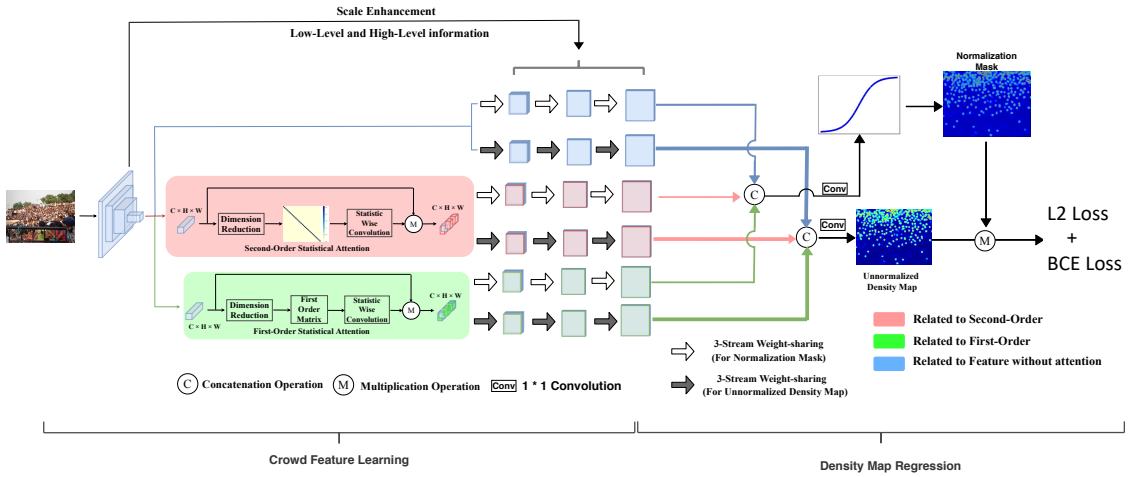


Fig. 3.1 The overall framework of SOFA-Net. Pink colored components are related to second-order statistics; Green colored components are related to the first-order statistics, Blue colored components are related to the feature from VGG16 backbone.

Based on $\mathbf{X}$, we can clearly see $c = \sum_{p \in \mathbf{X}} \mathbf{D}_p^{gt}$. At the inference stage, given model $\mathscr{F}$ and the query crowd image $\mathbf{X}'$, the density map $\mathbf{D}^{pr}$ and the corresponding counts c^{pr} can be calculated as follows:

$$\mathbf{D}^{pr} = \mathscr{F}(\mathbf{X}', \hat{\mathbf{W}}), \quad c^{pr} = \sum_{p \in \mathbf{X}} \mathbf{D}_p^{pr}, \tag{3.2}$$

where $\hat{\mathbf{W}}$ is the model parameters.

In previous works, CNN was the major technique for density map regression, yet these models tended to overestimate low density crowd or underestimate high density crowd [33]. To learn robust crowd feature, here we propose a deep attention network by exploring the second-order and first-order statistics, and aggregation of these two complementary information may be essential for reliable crowd density estimation. The structure of our method is shown in Fig. 3.1.

3.3.2 SOFA-Net: Second-Order and First-Order Attention Network

Fig. 3.1 shows the overall architecture of our SOFA-Net, which consists of crowd features learning part and the density map regression part.

The crowd feature learning part starts from a feature encoding backbone, where we use the first 13 layers in VGG16 network [192]. Our proposed second/first-order statistical attentions are computed on the VGG16 feature maps at the end of this backbone. Specifically,

the second-order statistics can be calculated based on (a derived) covariance matrix, while the first-order statistics can be extracted directly. The second-order and first-order statistics can be learned and transformed into attentions by a proposed Statistic-Wise Convolution operation. The two attentions can then be multiplied by crowd feature maps (extracted from VGG16), respectively for the second/first-order based features.

Given the second/first-order based features as well as the VGG16 features, we can estimate the crowd density map based on the generation blocks containing the Bilinear Up-sampling layer and the basic Convolution operations. Although the main goal is to learn the second/first-order statistics for crowd representation, we also fuse the VGG16 backbone feature because of the summarized high-level semantic representation [192]. The aggregation of these three feature types may improve the quality of the generated density map. Moreover, a normalization mask is also devised and learned to normalize the density map for better quality. It is worth noting that the normalization mask and the (unnormalized) density map are learned separately, which may make both matrices less correlated for better normalization effect. Finally, motivated by the effectiveness of U-Net [179], we concatenate the output low-level and high-level crowd features from different backbone layers, which also leads model robust for scale variation. The entire network is optimized by a Pixel-Wise L_2 loss and a Position-Wise Binary Cross Entropy (BCE) loss [125]. The details of these components are given the next subsections.

Second-Order Statistical Attention

Recent works [117, 31, 42, 230, 221] suggested effective representation can be learned by second-order statistics for deep convolution neural network. Here we propose to use second-order statistics as an important component in our SOFA-Net (pink rectangle in Fig. 3.1), which guides the model to learn the channel-wise spatial information for crowd. Given extracted feature maps (height(h), width(w), channel(c)) from VGG16 backbone $\mathbf{F} \in \mathbb{R}^{h\times w\times c}$, 1×1 convolution, ReLU function and Batch Normalization are applied to reduce the channel dimension from c to c' obtaining $\mathbf{F}' \in \mathbb{R}^{h\times w\times c'}$. Then $\mathbf{F}'$ is flattened into $\mathbf{Q} \in \mathbb{R}^{z\times c'}$ where $z = w \times h$. Covariance matrix $\mathbf{C}$ measuring the crowd correlation along channels can be formed as:

$$\mathbf{C} = \mathbf{Q}\bar{\mathbf{I}}\mathbf{Q}^T \tag{3.3}$$

where $\bar{\mathbf{I}} = \frac{1}{z}(\mathbf{Q} - \frac{1}{z}\mathbf{1})$ with $\mathbf{Q} \in \mathbb{R}^{c'\times c'}$ the identity matrix and $\mathbf{1} \in \mathbb{R}^{c'\times c'}$ is the matrix of all ones. After reshaping $\mathbf{C}$ to $\mathbf{C}' \in \mathbb{R}^{1\times c'\times c'}$, the Statistic-Wise Convolution (see Fig. 3.2) is devised to normalize the covariance to obtain the inherent feature correlation and transform

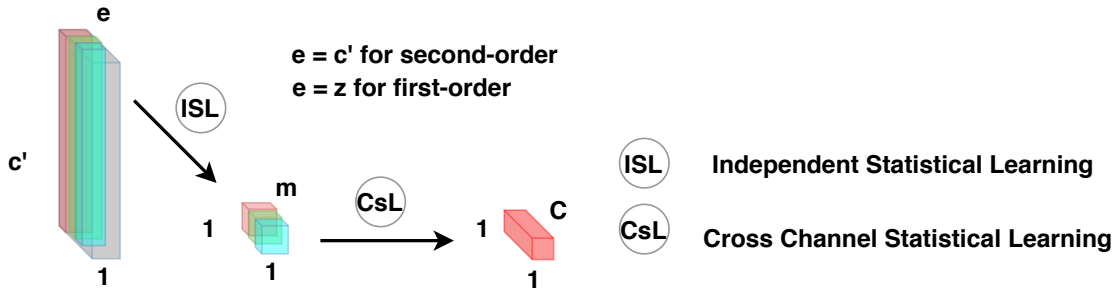


Fig. 3.2 Statistic-Wise Convolution including two components: ISL/CsL.

the second-order statistics into attention. The Statistic-Wise Convolution starts from a Independent Statistical Learning (see ISL component in Fig. 3.2) and the output feature maps $\mathbf{G} \in \mathbb{R}^{1\times1\times m}$ can be formed as:

$$\mathbf{G}_{1,1,m} = \mathbf{K}_{1,c',m} \circ \mathbf{C}'_{1,c',c'} \tag{3.4}$$

where $\circ$ denotes the element-wise multiplication. The $m(=c')$ convolution kernels $\mathbf{K}$ are applied to the c' channels with the same size $c' \times 1$ of feature maps (vector in this case). This operation learns the statistical dependency along channels. Then a 1×1 convolution (i.e., CsL component in Fig. 3.2) is applied to learn the statistics with sharing convolution kernels and increase the dimension to c, so the second-order statistics are transformed into second-order attention as $\mathbf{G} \rightarrow \mathbf{A}_{so}$. This attention $\mathbf{A}_{so}$ is multiplied by the feature $\mathbf{F}$ (extracted from VGG16) to refine the representation. Specifically, the crowd feature with second-order attention is for med as $\mathbf{F}_{so} = \mathbf{F} \bigotimes \mathbf{A}_{so}$, where $\bigotimes$ is the multiplication operation between corresponding feature maps.

First-Order Statistical Attention

First-order statistics have been widely adopted in many CNN-based classification tasks to guide the back-propagation [192, 171]. The first-order statistics were also known to capture the spatial characteristic for texture in images [42]. Here we utilize the first-order statistics to preserve the subjects' head edges with the discrimination of heads and non-heads. Following similar feature extraction (from VGG16, i.e., with $\mathbf{F} \in \mathbb{R}^{h\times w\times c}$) and dimension reduction procedures (i.e., with $\mathbf{F}' \in \mathbb{R}^{h\times w\times c'}$), we have the feature matrix $\mathbf{Q} \in \mathbb{R}^{z\times c'}$. Different from extracting second-order statistics, we directly use these spatial information. Specifically, after reshaping it into $\mathbf{Q}_{fo} \in \mathbb{R}^{1\times z\times c'}$, we learn the first-order attention $\mathbf{A}_{fo}$ by performing Statistic-Wise Convolution, i.e., $\mathbf{A}_{fo} = \mathbf{K}_{1\times z\times c'} \circ \mathbf{Q}_{fo}$. Similarly, the crowd feature with first-order attention is formed as $\mathbf{F}_{fo} = \mathbf{F} \bigotimes \mathbf{A}_{fo}$.

Density Map Estimation

The crowd density generation process are based on the aforementioned three feature types, i.e., crowd features $\mathbf{F}$ (e.g., extracted from VGG16), features with second-order statistical attention $\mathbf{F}_{so}$, and features with first-order statistical attention $\mathbf{F}_{fo}$. The final density map $\mathbf{D}^{pr}$ is generated based on two components, namely normalization mask $\mathbf{D}^{msk}$ and unnormalized density map $\mathbf{D}^{udm}$, as shown in Fig. 3.1. To reduce the correlation between these two components for better normalization effect, we train them through two-stream-like structure [269, 125].

A 3-stream weight-sharing scheme (via generation blocks f_G containing Bilinear Upsampling layer, 1×1 Convolution layer, 3×3 Convolution layer, Batch Normalization layer and ReLU activation layer) is used for each component. The unnormalized density map $\mathbf{D}^{udm}$ is generated as:

$$\mathbf{D}^{udm} = \{f_G(\mathbf{F}, \mathbf{W}^{udm}), f_G(\mathbf{F}_{so}, \mathbf{W}^{udm}), f_G(\mathbf{F}_{fo}, \mathbf{W}^{udm})\}. \tag{3.5}$$

Note $\mathbf{W}_{udm}$ are the shared weights among the three feature streams. Similarly, the normalization mask can be calculated via $\mathbf{D}^{msk} = f_{sigmoid}(f_{conv}(\mathbf{D}'))$, where

$$\mathbf{D}' = \{f_G(\mathbf{F}, \mathbf{W}^{msk}), f_G(\mathbf{F}_{so}, \mathbf{W}^{msk}), f_G(\mathbf{F}_{fo}, \mathbf{W}^{msk})\}, \tag{3.6}$$

and $\mathbf{W}_{msk}$ are the shared weights among the three feature streams. The final density map can be then estimated through $\mathbf{D}^{pr} = \mathbf{D}^{udm} \circ \mathbf{D}^{msk}$.

For better training effect [269], we also applied the scale enhancement strategies, which concatenate the different scale information from different layers in backbone to density map generation block. The scale information allows the model to be more robust for heads' scale variation.

3.3.3 Optimization

To learn the model parameters, we use two loss functions in SOFA-Net: pixel-wise loss and position-wise loss.

Pixel-Wise Loss The Pixel-wise loss L_2 is defined as:

$$L_2 = \frac{1}{z} \sum_{i=1}^{W} \sum_{j=1}^{H} (\mathbf{D}_{i,j}^{gt} - \mathbf{D}_{i,j}^{pr})^2 \tag{3.7}$$

where $\mathbf{D}^{pr}$ is the density map with height (H) and width (W) from the feed-forward operation and $z = W \times H$. It is the most widely used loss function on training deep convolution networks in crowd counting tasks [125, 128, 193].

Position-Wise Loss The position-wise loss L_{BCE} (Binary Cross Entropy) is calculated based on binarized ground truth $\mathbf{D}^b$, which comes from the $\mathbf{D}^{gt}$ based on a pre-defined threshold, and the (predicted) normalized density map (via Sigmoid) $\mathbf{S}^{pr} = f_{sigmoid}(\mathbf{D}^{pr})$ as follows:

$$L_{BCE} = -\frac{1}{z}\sum_{p=1}^{z}(\mathbf{D}_p^b \log(\mathbf{S}_p^{pr}) + (1-\mathbf{D}_p^b)\log(1-\mathbf{S}_p^{pr})) \tag{3.8}$$

The Position-wise loss enforces the learning process to discriminate the crowd locations for better quality of density map generation [125]. The final loss function for SOFA-Net optimization is formulated as (in this work, we assign the lambda equal to 0.9).

$$L = \lambda L_2 + (1-\lambda)L_{BCE} \tag{3.9}$$

Table 3.1 Performance comparison on four public crowd counting datasets

	UCF_QNRF		ShanghaiTech A		ShanghaiTech B		UCF_CC_50	
Method	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓
MCNN[259]	-	-	110.2	173.2	26.4	41.3	377.6	509.1
SCNN[183]	-	-	90.4	135.0	21.6	33.4	318.1	439.2
CSRNet[121]	-	-	68.2	115.0	10.6	16.0	266.1	397.5
RAZ-Net[125]	116	195.0	65.1	106.7	8.40	14.1	-	-
L2SM[236]	104.7	173.6	64.2	98.40	7.20	11.1	188.4	315.3
DSSINet[128]	99.1	159.2	60.63	96.04	6.85	10.34	216.9	302.4
MBTTBF[193]	97.5	165.2	60.2	94.10	8.00	15.5	233.1	300.9
SPANet[33]	-	-	59.4	92.50	**6.50**	**9.9**	232.6	311.7
Ours	**96.2**	**158.7**	**57.5**	**92.12**	6.80	10.38	**185**	**281**

3.4 Experiment

3.4.1 Implementation Details

Network Setting The first 13 VGG16 layers(pre-trained model on ImageNet) were used to initialize the corresponding layers in SOFA-Net. Other parameters were initialized by Gaussian disstribution with zero mean and 0.01 standard deviation. We set batch size to 50 and epoch number to 2000 in our experiments. We performed bilinear interpolation for any images less than 512× 512, and the images were fed into network after randomly

being cropped to 400 $\times$ 400 pixels. Observing that the images were collected from various situations of illumination, we adjusted the images by gamma contrast [0.5, 1.0] with the probability 25%. There are a few gray images in some datasets (e.g., ShanghaiTech_A), so in data augmentation we randomly converted a few (10%) for robust model training.

Datasets Our method was evaluated on the four popular public datasets, i.e., UCF_QNRF [88], ShanghaiTech [259](Part A and B), and UCF_CC_50 [87]. Out of them, UCF_QNRF contains large density variations and the subject number ranges from 49 to 12865. UCF_CC_50 includes extreme crowd scenes with serious noise. ShanghaiTech part A is very congested with noise, while ShanghaiTech part B is not congested. Following the protocol used in [125, 269], we generated the ground truth by a fixed Gaussian kernel. Also, the ground truth binary maps were generated by setting the threshold to 0.001 based on the ground truth density maps[125, 269]. For most train/test configurations, we followed the default protocols in the original papers (i.e., UCF_QNRF [88], UCF_CC_50 [87], ShanghaiTech_AB [259]). Notably, due to the limited sample numbers in UCF_CC_50, following [87] we set 5-fold cross validation for evaluation.

3.4.2 Evaluation Metrics

Following most existing works, the Mean Absolute Error (MAE) and Mean Square Error(MSE) were used as the evaluation metrics. Given predicted subject number c^{pr} (which can be inferred from $\mathbf{D}^{pr}$, see Eq. (2)), for N test crowd images the evaluation metrics can be calculated by:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|c_i^{pr} - c_i^{gt}|, MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(c_i^{pr} - c_i^{gt})^2} \tag{3.10}$$

where c^{gt} denotes the ground truth heads' counting number.

3.4.3 Experimental Results

Model Comparison Table 3.1 shows the results of our SOFA-Net and other state-of-the-arts on four afore-mentioned datasets. Our SOFA-Net outperforms others on most of the datasets (except ShanghaiTech Part B), which suggests its effectiveness on general crowd modelling tasks. Compared with the most recent works (i.e., methods in 2019) on UCF_QNRF dataset, SOFA-Net reaches much better results with further error reduction (i.e., in terms of MAE 1.3 - 19.8 and MSE 0.5 - 36.3) than other methods. On ShanghaiTech Part A, our method is also much better in terms of both MAE and MSE. For ShanghaiTech Part B which was collected from shopping street with less crowded scenes, it can be seen that all the methods have good

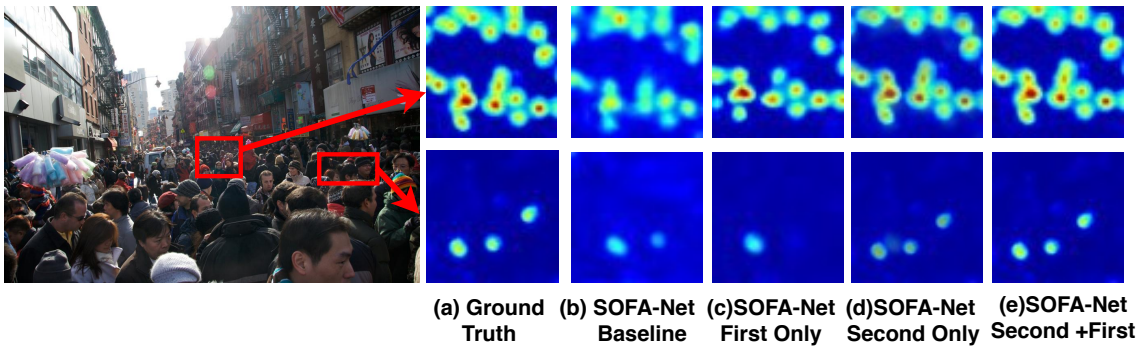


Fig. 3.3 The generated maps based on different settings of SOFA-Net in high density area (Top) and low density area (Bottom).

results in this relatively sparse and simple dataset. The performance can be further boosted by fusing other complementary information via the ensemble learning [65] or multi-stream structure [34, 33]. Nevertheless, our method outperforms most of the algorithms, and is comparable with state-of-the-art. UCF_CC_50 dataset, which includes very crowded scenes with high-levels of noises, was considered as the most challenging dataset. We can see our SOFA-Net can model the crowd counting tasks in these extreme conditions well, with much lower errors than other works (i.e., 3.4-234.5 in MAE and 19.9-260.6 in MSE).

Qualitative Analysis To understand better the effect of the proposed second/first-order statistical features, we visualized a challenging crowd image and the generated density maps (under different settings) in Fig. 3.3, from which some interesting observations can be made:

- from Fig. 3.3b, we can see without second/first-order statistical features (i.e., with feature $\mathbf{F}$ only), the generated crowd map is very blurry.
- with first-order statistical features (Fig. 3.3c, i.e., with features $\mathbf{F}$,$\mathbf{F}_{fo}$), we can see clear boundaries among heads as enhanced discrimination, yet it cannot model the center-likelihood of dense heads areas well (with relatively low likelihoods in the centers).
- Retaining the selectivity of spatial information, features with second-order statistical attention (Fig. 3.3d, i.e., with features $\mathbf{F}$,$\mathbf{F}_{so}$) can well preserve heads' areas (with high and precise likelihoods), which leads to the accurate counting. Finally aggregating both features can yield precise estimation (Fig. 3.3e, i.e., with features $\mathbf{F}$,$\mathbf{F}_{fo}$,$\mathbf{F}_{so}$).

From Fig. 3.3, we can clearly see second/first-order information are complementary for better crowd density map generation. We also generated several density maps in some challenging scenarios (as shown in Fig. 3.4), and results suggested its effectiveness even when there were more than thousand of people in crowd scenes.

Fig. 3.4 Some density maps generated by SOFA-Net; From top to bottom: original images, ground truth maps and generated maps

3.4.4 Ablation Study

We also conducted ablation studies to quantitatively assess the core components in our SOFA-Net. ShanghaiTech part A, which covers various subject number in different scenes, was used as the benchmark dataset.

ShanghaiTech Part A dataset		
SOFA-Net	MAE	MSE
Convolution	59.8	96.3
Statistic-Wise Convolution	**57.5**	**92.1**

Table 3.2 on the effect of Statistics-Wise Convolution

ShanghaiTech Part A dataset			
SOFA-Net	Features	MAE	MSE
No Attention	$\mathbf{F}$	68.6	109.3
First-Order	$\mathbf{F} + \mathbf{F}_{fo}$	65.6	104.2
Second-Order	$\mathbf{F} + \mathbf{F}_{so}$	60.8	97.1
Second/First-Order	$\mathbf{F}+\mathbf{F}_{fo}+\mathbf{F}_{so}$	**57.5**	**92.1**

Table 3.3 on the effect of second/first-order statistical attentions

Effect on Second/First-Order Statistical Attention The core contribution of this work is the proposed second/first-order statistical attentions for robust representation learning. Fig. 3.3 demonstrated the effect of both components in a qualitative manner, and here we study them quantitatively. In Table 3.2, we report the SOFA-Net's results under different settings. With different attention types, we can use the corresponding feature combinations (e.g., $\mathbf{F}$,$\mathbf{F}_{fo}$,$\mathbf{F}_{so}$) to generate the density maps for crowd counting. We can clearly see that second-order statistical attention contributes the most to the performance, and the error rate can be further reduced if the complementary second/first-order statistical features were aggregated.

Effect on Statistic-Wise Convolution Statistic-Wise Convolution is a tailored operation that is proposed to learn second/first-order statistical attentions. The results in Table 3.3 suggests its effectiveness when compared with the standard convolution operation.

ShanghaiTech Part A dataset		
SOFA-Net	MAE	MSE
Without (Normalization) Mask	62.1	98.1
Mask (full-weight-sharing)	61.3	97.4
Mask (3-stream-weight-sharing)	**57.5**	**92.1**

Table 3.4 on the effect of normalisation masks

Effect on Normalization Mask In this work, we also trained a normalization mask to scale the generated density maps to avoid trivial results mostly in non-heads areas. In Table 3.4, we compared SOFA-Nets with/without normalization masks. For models with normalization masks, we also reported results with two different training strategies, i.e., the proposed 3-stream-weight-sharing scheme, as well as the full-weight-sharing scheme. Specifically, the former shared weights among the three feature streams $\mathbf{F}$,$\mathbf{F}_{fo}$,$\mathbf{F}_{so}$ to learn $\mathbf{W}^{udm}$ in Eq.(5) and $\mathbf{W}^{msk}$ in Eq.(6), respectively, while the latter shared weights among the three feature streams as well as the two tasks (with $\mathbf{W}^{udm} = \mathbf{W}^{msk}$). From Table 3.4, we can clearly see the normalization mask trained by the proposed 3-stream-weight-sharing scheme is much better than other structures. The result also suggests that a less correlated normalization mask (e.g., trained without weight-sharing between tasks) may further reduce the errors in such density map regression tasks.

3.5 Conclusion

In this work, we proposed SOFA-Net, which can extract second/first-order statistical attentions to learn robust representations for reliable crowd density map regression. The

experimental results suggested second/first-order based features are complementary, and aggregating both features is feasible to reduce the error rate substantially for crowd density estimation. Also, the proposed method outperformed other state-of-the-arts in most (challenging) datasets. Although additional experiments and theoretical findings are necessary to draw the final conclusions on the benefit of applying second/first-order statistics for crowd counting, this work empirically demonstrates a simple yet effective way on modelling the crowd in challenging scenarios.

Chapter 4

Attention-based Feature Integration for Semi-Supervised Crowd Counting

Chapter 3 indicates that our proposed attention mechanism is effective in supervised learning with annotated data, while in this chapter we aim to further evaluate how the proposed attention mechanism works with the large amount of unlabelled data in a semi-supervised setting. Automatic Crowd behavior analysis can be applied to effectively help the daily transportation statistics and planning, which helps the smart city construction. As one of the most important keys, crowd counting has drawn increasing attention. Recent works achieved promising performance but relied on the supervised paradigm with expensive crowd annotations. To alleviate the annotation cost in real-world transportation scenarios, in this work we proposed a semi-supervised learning framework S^4*Crowd*, which can leverage both unlabeled/labeled data for robust crowd counting. In the unsupervised pathway, two self-supervised losses were proposed to simulate the crowd variations such as scale, illumination, etc., based on which and the supervised information pseudo labels were generated and gradually refined. We also proposed a crowd-driven recurrent unit Gated-Crowd-Recurrent-Unit (GCRU), which can preserve discriminant crowd information by extracting second-order statistics, yielding pseudo labels with improved quality. A joint loss including both unsupervised/supervised information was proposed, and a dynamic weighting strategy was employed to balance the importance of the unsupervised loss and supervised loss at different training stages. We conducted extensive experiments on four popular crowd counting datasets in semi-supervised settings. Experimental results supported the effectiveness of each proposed component in our S^4Crowd framework. Our method also outperformed other state-of-the-art semi-supervised learning approaches on these crowd datasets.

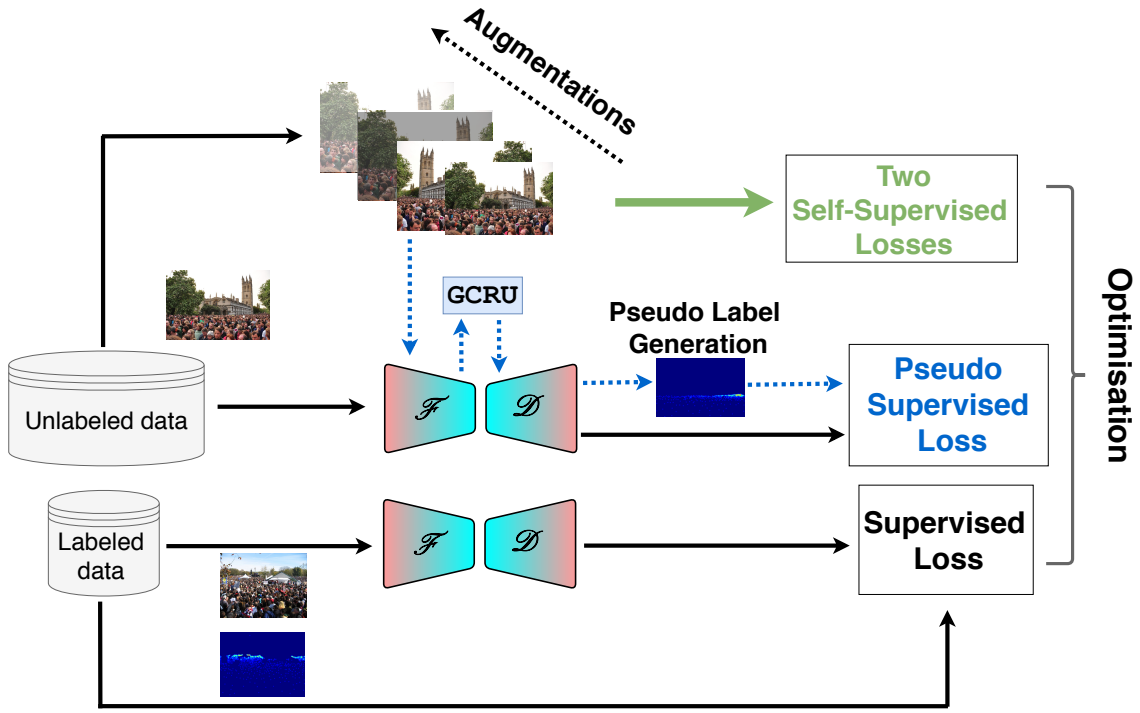


Fig. 4.1 The Proposed S^4Crowd Framework

4.1 Introduction

Crowd counting plays a vital role in real-world applications, specifically, it has had obviously increasing impact for intelligent transportation in smart cities [219, 126, 47, 247]. Automatic Crowd counting can directly help the human operators to dynamically and accurately obtain the real-time statistics of objects (i.e., human traffic) traffic or the level of congestion in the various conditions of transportation. With the help of the crowd counting, transportation planning can be effectively adjusted and some traffic problems can be quickly solved or avoided.

In this work, we focus on image-based crowd scenes (captured from public surveillance) estimation with explicit pedestrians counting. The ways of obtaining the crowd statistics have been progressively explored from detection-based [213], segmentation-based approaches [262] to head density estimation based methods [115]. Compared with detection/segmentation-based approaches, density-based methods tended to be less sensitive to occlusions in dense crowd, and they became the mainstream approaches nowadays.

For density-based methods, Convolution Neural Network (CNN) was the major technique [83, 93] for crowd representation learning and density estimation. While existing CNN-density-based crowd counting methods achieved promising performance, they relied heavily

on labeled data, and the acquisition of labels can be labour-intensive and time-consuming. For example, it took 2000 human-hours to annotate 1535 images (with 1251642 locations of crowd heads) in the UCF_QNRF dataset [88]. When with inadequate annotated labels, overfitting may occur.

To reduce the overfitting effect and the annotation costs [219, 126, 47, 247], semi-supervised learning, which can take advantage of both labelled/unlabelled data, has been widely used in many computer vision tasks such as semantic segmentation [149], object detection [261], action recognition [190]. Most recently, semi-supervised learning was also employed in the crowd counting community. Sindagi et al. proposed a Gaussian-Process iterative learning to estimate the pseudo labels for unlabeled crowd data [194], and Liu et al. proposed to learn the crowd representation from unlabeled data with some surrogate segmentation tasks [134]. Both works[194, 134] demonstrated the effectiveness of leveraging the unlabelled crowd data, yet their semi-supervised schemes failed in modelling the variations of crowds (e.g., head scale, illumination, perspective distortion, etc.), which were considered as indispensable factors for reliable crowd modelling.

To address this issue, we proposed to model different crowd variations in a self-supervised manner, specifically with two regularization terms. In the meantime, the pseudo labels of the un-annotated data were generated and gradually refined for better model generalization. In our semi-supervised scheme, a joint cost function including both supervised/unsupervised loss was proposed, and a dynamic weighting strategy was employed to balance both losses at different training stages.

Specifically in the unsupervised pathway, a sequence of augmentation functions (i.e., image transformations) were applied to each unlabeled image for simulating different variations. To better encode the high-order-tensor augmented image sequences for high-quality pseudo labels generation, we proposed a novel recurrent unit, namely Gated-Crowd-Recurrent-Unit (GCRU). Compared with the conventional GRU[35], GCRU further considered the second-order statistics of crowd, which was demonstrated as informative features in crowd modelling [52]. The general structure of our proposed S^4Crowd is shown in Fig.4.1. In this work, our main contribution can be summarised as:

- We proposed a novel semi-supervised crowd counting framework S^4Crowd, which included several key components (variations modelling, pseudo labeling, dynamic weight training scheme, etc.) for robust crowd modelling when with limited labeled data.

- We proposed two self-supervised losses/regularisation terms, i.e., Crowd Scale Equivariance (CSE), Crowd Entropy Consistency (CEC), based on which we modelled the crowd variations in an unsupervised manner.
- A crowd-driven recurrent unit, i.e., Gated-Crowd-Recurrent-Unit (GCRU) was proposed to encode high-order-tensor crowd sequences, which can also learn the informative second-order crowd statistics for better crowd modelling.
- Extensive experiments were conducted, and the proposed components in our semi-supervised S^4Crowd framework were well studied. Our method outperformed other semi-supervised crowd counting algorithms in various settings.

4.2 Related Work

4.2.1 Crowd Counting

Lempitsky and Zisserman casted the crowd counting problem as the efficient density estimation of only head area [115], and the integrating over density map gives the counting of crowd. Furthermore, CNN has been the main techniques for density-based crowd counting [33, 83, 251] to learn the crowd representation and estimate the density maps. In order to alleviate additional annotation cost for crowd model training, semi-supervised crowd counting [134, 194, 263] has drawn some attentions. Specifically, a Gaussian-Process (GP) was used to model the relationship between the feature latent space and ground truth [194] in labeled data, and generated pseudo labels for unlabeled data. In [134], Liu et al. proposed a self-training framework with surrogate segmentation tasks, and the binary inter-relationship of crowd and non-crowd leads to the accurate density prediction. Also, Zhao et al. proposed a active learning framework [263] to select a small fraction of most informative data to be annotated to train the model.CLRNet [49] is a model applied to video crowd counting. It utilizes spatial and temporal information by adding a cross-position relationship module to learn complex spatiotemporal relationships between different frames in a video sequence. This network includes two core modules: a backbone network for feature extraction and a cross-local relationship module for temporal feature integration. Additionally, it employs a density-aware loss function to more effectively handle density variations in crowd counting. CrowdFormer [240] is a crowd counting method based on the visual transformer structure, which uses overlap patching techniques. This method copes with the spatial variation of crowd density through a patching strategy of the input image. CrowdFormer also integrates global and local information of image patches through a multi-level feature aggregation

method. It uses position encoding and spatial attention mechanisms to process input image patches. Overlap patching technology reduces noise in density maps and improves counting accuracy. MAN [124] generates a set of attention maps for each scale based on the multi-scale features of the input image. The method combines these attention maps using a multifaceted attention mechanism to highlight regions closely related to crowd counting. It also introduces a feature fusion module to fuse multi-scale features in a weighted manner. DMCNet [218] designed a two-stage architecture. The first stage is a feature extraction network based on the pre-trained VGG16 network. The second stage is a dynamic hybrid counting network, which contains a series of counting modules with different receptive field sizes. The receptive field size of these modules is adaptively adjusted according to the size of the input image, achieving position-independent crowd counting. The network combines the predictions of each counting module through a dynamic weighting scheme. The crowd counting task is limited by data annotation. The currently available labeled data is not only limited in quantity, but also of a single type. Additionally, the process of collecting and labeling data for a specific task is time-consuming and laborious. Therefore, the research field of crowd counting needs to develop a model that can utilize unlabeled data to expand the data set. Sam et al. proposed a semi-supervised learning method for crowd counting [182], which mainly trained the model through unlabeled data.

4.2.2 Semi-Supervised Learning

Semi-supervised learning aims to take advantage of both labeled and unlabeled data for robust model training. To obtain the useful information from unlabeled data [257], data augmentation (e.g., flip, greayscale, rotation, etc.) has been the crucial operation for semi-supervised learning, which is used to explore the consistency among the different transformations of input space [12, 224, 64, 137, 221]. The consistency [12, 116] has been widely derived as regularization terms to constrain the model prediction consistently cross the different data augmentation on same target. Moreover, Pseudo-Labeling is another vital semi-supervised approach [113, 234], where the pseudo labels are generated for unlabeled data to supervise the model training. The rationale behind is that different augmented versions of the same target should closely lie in the feature space. Every day even every second, there is a massive volume of data streamed from traffic monitoring or public surveillance and it may not be possible to annotate it. Therefore, successfully applying semi-supervised learning is valuable, which can improve the capability of intelligent transportation systems via not only existing limited labelled data but also the lots of unlabelled data.The self-training algorithm generates pseudo labels by leveraging the model's prediction confidence on unlabeled data. The entropy minimization strategy, as outlined in [61], is designed to prevent the model's decision

boundary from intersecting high-density data regions. It achieves this by encouraging the model to make low-entropy predictions for unlabeled data, which are then added to the labeled dataset. This forms a part of implementing semi-supervised learning (SSL) within a standard supervised learning framework. The pseudo-label method, referenced in [114], offers a straightforward yet effective means for training semi-supervised neural networks. It utilizes both labeled and unlabeled data in supervised learning scenarios. In this method, the model initially undergoes training on labeled data through standard supervision techniques and employs a cross-entropy loss function. For unlabeled data, the model predicts a set of samples and selects the prediction with the highest confidence as the pseudo label - essentially, the label with the highest prediction probability. Noisy Student, introduced in [234], is a semi-supervised learning technique inspired by knowledge distillation, discussed in [78]. It involves using student models of the same or greater size. The process begins with training an EfficientNet model, as mentioned in [202], on labeled images. This model then generates pseudo labels for unlabeled data. Subsequently, a larger EfficientNet model is trained as a student, utilizing a mix of labeled and pseudo-labeled data. RandAugment, cited in [40], is applied for data augmentation during training, and the student model incorporates dropout and random depth strategies. Meta Pseudo Labeling (MPL), discussed in [165], entails a teacher model assigning probability distributions to input samples to facilitate the training of the student model. Throughout the training, the teacher model monitors the student's performance on a separate validation set. This monitoring helps the teacher model learn to produce an optimal target distribution, aiding the student model in achieving strong validation performance. Semi-supervised learning is valuable in crowd counting primarily due to the scarcity of labeled data, which is costly and time-consuming to obtain. It enables the use of a limited set of labeled images alongside a larger set of unlabeled images, making the training process more efficient and cost-effective. Additionally, it helps in capturing the wide variability in crowd scenes and improves the model's generalization ability, making it more robust to different and unseen scenarios.

4.3 The Proposed Method

4.3.1 Problem Statement

Density-based crowd counting has been regarded as a regression problem [83, 115], and pixel-wise density should be estimated before counting. Given a crowd image sample $\mathbf{x}$ with e pedestrians' heads $\mathbf{V} = \{\mathbf{v}_q \in \mathbb{N}^2\}_{q=1}^{e}$, where $\mathbf{v}_q$ is the x-y coordinate of the q_{th} head's center point, and the ground truth density map of $\mathbf{x}$ can be constructed by e Gaussian distributions

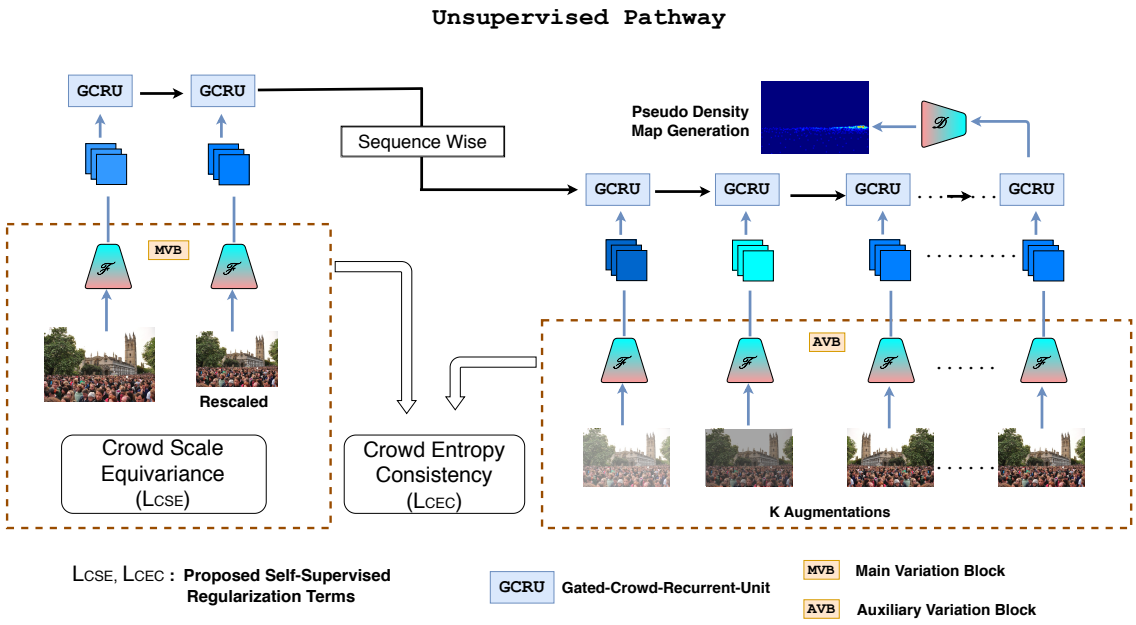


Fig. 4.2 The overall architecture of the proposed Unsupervised pathway, which consists of CSE/CEC regularization terms and pseudo labels generation with Gated-Crowd-Recurrent-Unit(GCRU).

($\mathcal{N}$) with $\mathbf{V}$ as the mean values. Then the crowd can be counted by calculating the the integral of the density map. The ground truth density map $\mathbf{y}^{gt}$ can be calculated in the same way in Equation 3.1.

where $p \in \mathbf{x}$ denotes the image pixels, σ is a fixed [93] value with a very small number spanning a few pixels [115], and the head counting can be calculated via $e = \sum_{p \in \mathbf{x}} \mathbf{y}_p^{gt}$.

A typical crowd density estimation model normally included a crowd representation learning module $\mathscr{F}$ and a density regression module $\mathscr{D}$ [122, 93]. Given the predicted density map $\mathbf{y}^{pr} = \mathscr{D}(\mathscr{F}(\mathbf{x}))$, the supervised training process is to minimise the following MSE loss [1]

$$\mathscr{L}_S = \frac{1}{WH} \sum_{w=1}^{W} \sum_{h=1}^{H} \left(\mathbf{y}_{w,h}^{pr} - \mathbf{y}_{w,h}^{gt} \right)^2 \tag{4.1}$$

where H and W are image height and width respectively. In the inference stage, given the trained model and a query crowd image $\mathbf{x}'$, the predicted density map $\mathbf{y}^{pr}$ and the corresponding counts e^{pr} can be calculated as follows:

$$e^{pr} = \sum_{p \in \mathbf{x}'} \mathbf{y}_p^{pr}, \text{where } \mathbf{y}^{pr} = \mathscr{D}(\mathscr{F}(\mathbf{x}')). \tag{4.2}$$

[1] Note for demonstration, we only used sample-wise loss here

However, the aforementioned supervised pathway may require costly annotations. When with inadequate labelled data, overfitting may occur, yielding unreliable $\mathscr{F}$ and $\mathscr{D}$ for crowd modelling.

There are several ways to alleviate the overfitting effect without additional annotation requirement, such as regularisation, unsupervised learning (e.g., self-supervised learning), semi-supervised learning, etc. For reliable $\mathscr{F}$ and $\mathscr{D}$ estimation, in this work we proposed a semi-supervised framework, where both unsupervised and supervised losses were combined for model training. Specifically, in the unsupervised learning pathway two self-supervised losses were derived to model the crowd variations, based on which a crowd-driven recurrent unit GCRU was proposed to better encode crowd sequential information for pseudo label generation. Based on the pseudo labels, the trained $\mathscr{F}$ and $\mathscr{D}$ can be more robust to crowd variations, yielding more reliable crowd density maps.

In the semi-supervised settings, the training dataset was split into two subsets, namely the labeled set $\mathscr{X}^l = \{\mathbf{x}_n^l, \mathbf{y}_n^l\}_{n=1}^{N^l}$ and the unlabeled set $\mathscr{X}^u = \{\mathbf{x}_m^u\}_{m=1}^{N^u}$, where N^l is sample number of labeled data and N^u is sample number of unlabeled data.

4.3.2 Unsupervised Pathway

In Fig.4.2, we demonstrate the general idea of unsupervised pathway in our semi-supervised framework. Here we aim at generating pseudo labels for the unlabeled set $\mathscr{X}^u$ such that the trained representation can be less sensitive to some variations. Two self-supervised regularization terms (i.e., CSE and CEC) were proposed, based on which an augmentation sequence can be generated to simulate the crowd variations, before applying GCRU for pseudo label generation.

Crowd Scale Equivariance (CSE) Modelling variations such as head scale [83, 139] can be an effective way for improved crowd density estimation, yet both works relied heavily on labelled data. Motivated by this, we also modelled the head scale variation, but in an unsupervised manner. The lack of labels makes it a challenging modelling task and here we proposed a loss (or regularisation term), named Crowd Scale Equivariance (CSE), based on self-supervised learning concept. The rationale behind is that the crowd density distribution (and the counting result) should not be affected by different head/image scales. Given unlabeled crowd image sample $\mathbf{x}_m^u$, and a re-scaling transformation function $\mathscr{R}$ (with re-scaling rate r), the CSE loss can be defined as

$$\mathscr{L}_{CSE} = \|\mathscr{D}(\mathscr{F}(\mathscr{R}(\mathbf{x}_m^u))) - \mathscr{R}(\mathscr{D}(\mathscr{F}(\mathbf{x}_m^u)))\|_1 \tag{4.3}$$

It is worth noting that both $\mathscr{F}(\cdot)$ and $\mathscr{D}(\cdot)$ do not include dense layers and they are flexible for different input image sizes. We used L_1-norm because it is less sensitive to outliers [154]. This CSE regularisation is demonstrated in the Main Variation Block (MVB) in Fig. 4.2, which also outputs a sequence of feature maps $\mathbf{G}_{MVB}^{\mathbf{x}_m^u}$ for future processing:

$$\mathbf{G}_{MVB}^{\mathbf{x}_m^u} = \{\mathscr{U}(\mathscr{F}(\mathscr{R}(\boldsymbol{x}_m^u))), \mathscr{F}(\boldsymbol{x}_m^u)\}, \tag{4.4}$$

where $\mathscr{U}(\cdot)$ is an up-sampling function.

Crowd Entropy Consistency (CEC) Consistency regularisation is a popular notion in semi-supervised learning for robust model training [212, 116], and the rationale behind is that the target image's semantics should keep unchanged during image transformations. It is worth noting that CSE is a special case when the transformation is image scaling.

For robust crowd modelling, we further extended this semantics-invariant concept to other transformations to reduce the effect of crowd variations such as illumination, image quality. As shown in Fig. 4.2, the Auxiliary Variation Block (AVB) includes K weakly augmented crowd images via image operations $\{\mathscr{A}_k(\cdot)\}_{k=1}^K$ such as illumination adjustment, grayscale conversion, gamma adjustment, etc. For crowd image $\boldsymbol{x}_m^u$, based on $\{\mathscr{A}_k(\cdot)\}_{k=1}^K$ and $\mathscr{F}(\cdot)$, AVB outputs a sequence of feature maps $\mathbf{G}_{AVB}^{\mathbf{x}_m^u}$:

$$\mathbf{G}_{AVB}^{\mathbf{x}_m^u} = \{\mathscr{F}(\mathscr{A}_1(\boldsymbol{x}_m^u)), \mathscr{F}(\mathscr{A}_2(\boldsymbol{x}_m^u)), .., \mathscr{F}(\mathscr{A}_K(\boldsymbol{x}_m^u))\}. \tag{4.5}$$

Given the output sequences $\mathbf{G}_{MVB}^{\mathbf{x}_m^u}$ and $\mathbf{G}_{AVB}^{\mathbf{x}_m^u}$, we further calculated the sequence expectations $\bar{\mathbf{G}}_{MVB}$ and $\bar{\mathbf{G}}_{AVB}$. Motivated by [161], which used entropy maps to address domain gap in unsupervised semantic segmentation tasks, we proposed the Crowd Entropy Consistency (CEC) regularisation/loss:

$$\mathscr{L}_{CEC} = \|\mathbf{E}_{MVB} - \mathbf{E}_{AVB}\|_1, \tag{4.6}$$

where $\mathbf{E}_{MVB}$ and $\mathbf{E}_{AVB}$ are the pseudo entropy maps defined as follows:

$$\begin{aligned} \mathbf{E}_{MVB} &= \frac{-1}{\log C} \sum_{c \in C} \mathbf{P}_{MVB}^c \log \mathbf{P}_{MVB}^c, \\ \mathbf{E}_{AVB} &= \frac{-1}{\log C} \sum_{c \in C} \mathbf{P}_{AVB}^c \log \mathbf{P}_{AVB}^c. \end{aligned} \tag{4.7}$$

In Eq (4.7), $\mathbf{P}_{MVB}^c = \sigma(\bar{\mathbf{G}}_{MVB}^c)$ and $\mathbf{P}_{AVB}^c = \sigma(\bar{\mathbf{G}}_{AVB}^c)$ where $\sigma(\cdot)$ is the sigmoid function; C is the feature map number, and $\log C$ is a normalisation term.

The motivation of $\mathscr{L}_{CEC}$ is from the information gain concept in information theory. For decision tree, one of the tree-splitting principle is to select the most discriminant feature with the maximal information gain (measured in terms of entropy differences). However, in our case, there should not be any significant entropy differences (i.e., information gain) between two different types of transformed images (via $\mathscr{R}(\cdot)$ and $\{\mathscr{A}_k(\cdot)\}_{k=1}^{K}$) in the "variation-insensitive" feature space. In this case, $\mathscr{L}_{CEC}$ can be used as a regularisation term to minimise the impact of crowd variations.

Gated-Crowd-Recurrent-Unit For an unlabeled image $\mathbf{x}_m^u$, we generated the pseudo label for better training effect. Given sequence $\mathbf{G}(\mathbf{x}_m^u) = \{\mathbf{G}_{MVB}^{\mathbf{x}_m^u}, \mathbf{G}_{AVB}^{\mathbf{x}_m^u}\}$, consisting of the outputs from MVB and AVB in Fig. 4.2, we propose Gated-Crowd-Recurrent-Unit (GCRU) for sequence encoding. The internal structure of GCRU can be found at Fig. 4.3. Compared with standard deep model for sequential data (e.g., GRU [35]), our GCRU not only can preserve the high-order tensor structure, but also can extract the crowd-driven features for modelling. For simplicity, for the rest of this subsection we used $\mathbf{G}$ instead of $\mathbf{G}(\mathbf{x}_m^u)$.

GRU [35] is a popular sequential modelling method, yet it takes vector sequence as input while in our case, the input is high-order tensor sequence $\mathbf{G} = \{\mathbf{G}_t \in \mathbb{R}^{C\times H\times W}\}_{t=1}^{K+2}$. Our GCRU can preserve the tensor structure of the crowd data, and given current input $\mathbf{G}_t$ and previous state $\mathbf{H}_{t-1}$ it can calculate the current high-order hidden state $\mathbf{H}_t \in \mathbb{R}^{C\times H\times W}$

$$\mathbf{H}_t := \text{GCRU}(\mathbf{G}_t, \mathbf{H}_{t-1}),\ t \in \{1, 2.., K+2\}. \tag{4.8}$$

Similar to GRU, our GCRU also has two gates, namely the reset gate ($\mathbf{Q}^{rs} \in \mathbb{R}^{C\times H\times W}$) and the update gate ($\mathbf{Q}^{ud} \in \mathbb{R}^{C\times H\times W}$). The reset gate is used to reduce noises in $\mathbf{G}_t$, and refined signal $\hat{\mathbf{G}}_t \in \mathbb{R}^{C\times H\times W}$ can be calculated via:

$$\hat{\mathbf{G}}_t := \mathbf{Q}^{rs} \odot \mathbf{G}_t, \text{where } \mathbf{Q}^{rs} := \sigma(\text{Conv}([\mathbf{G}_t \oplus \mathbf{H}_{t-1}])), \tag{4.9}$$

where $\odot$ is the entry-wise multiplication; $\oplus$ is concatenation; $\text{Conv}(\cdot)$ is a convolutional operation for dimensionality reduction and $\sigma(\cdot)$ denotes the sigmoid function. Then the update gate can be applied between $\hat{\mathbf{G}}_t$ and $\mathbf{H}_{t-1}$.

In the most recent work [52], the authors demonstrated the effectiveness of second-order statistics in crowd modelling. They calculated the channel-wise covariance matrix, which was converted into correlation maps on learning the crowd representation. Motivated by this, we proposed the crowd-driven update gate by employing the crowd second-order statistics. In our case, based on a two-stream structure, given $\mathbf{U} = \hat{\mathbf{G}}_t \oplus \mathbf{H}_{t-1}$ we extracted both the channel-wise correlation maps $\mathbf{U}_c \in \mathbb{R}^{C\times 1\times 1}$ [52] and spatial-wise correlation maps

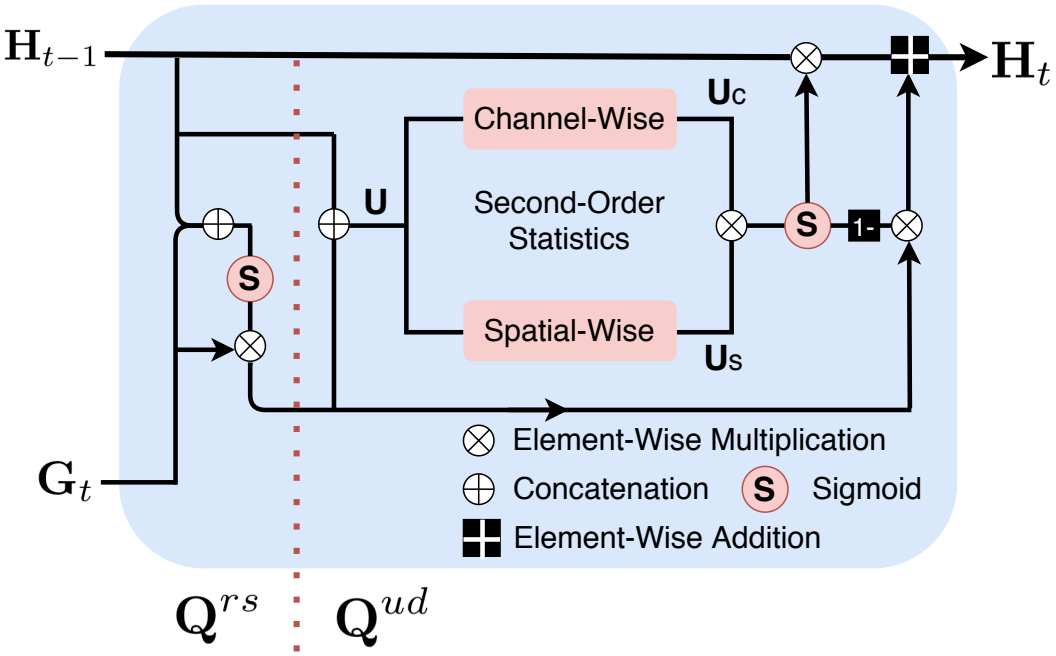


Fig. 4.3 Gated-Crowd-Recurrent-Unit

$\mathbf{U}_s \in \mathbb{R}^{1\times Z\times Z}$ respectively, where $Z = WH$. Given that, the current hidden state $\mathbf{H}_t$ can be calculated as follows:

$$\begin{aligned} &\mathbf{H}_t := \mathbf{Q}^{ud} \odot \mathbf{H}_{t-1} + (1 - \mathbf{Q}^{ud}) \odot \hat{\mathbf{G}}_t, \\ &\text{where } \mathbf{Q}^{ud} := \sigma(\text{Conv}(\mathbf{U}_c * \mathbf{U}_s)), \end{aligned} \tag{4.10}$$

where $*$ is the tensor-multiplication.

Pseudo Label Generation: GCRU's output at final step can be used for pseudo label generation. For unlabelled image $\mathbf{x}^u_m$, the pseudo label $\mathbf{y}^u_m$ can be written as:

$$\mathbf{y}^u_m = \mathscr{D}(\text{GCRU}(\mathbf{G}(\mathbf{x}^u_m)_{K+2})), \tag{4.11}$$

based on which the pseudo supervised (PS) loss $\mathscr{L}_{PS}$ can be calculated via:

$$\mathscr{L}_{PS} = \frac{1}{WH} \sum_{w=1}^{W} \sum_{h=1}^{H} \left(\mathscr{D}(\mathscr{F}(\mathbf{x}^u_m))_{w,h} - \mathbf{y}^u_{m,w,h} \right)^2. \tag{4.12}$$

4.3.3 Training Strategy

With supervised loss $\mathscr{L}_S$ (as defined in Eq. (4.1)) and unsupervised losses $\mathscr{L}_{CSE}, \mathscr{L}_{CEC}, \mathscr{L}_{PS}$ (as defined in Eq. (4.3), Eq. (4.6), Eq. (4.12) respectively), we can perform semi-supervised learning. Given labeled set $\mathscr{X}^l = \{\mathbf{x}^l_n, \mathbf{y}^l_n\}_{n=1}^{N^l}$ and unlabeled set $\mathscr{X}^u = \{\mathbf{x}^u_m\}_{m=1}^{N^u}$, we con-

structed the following joint loss function to be minimised:

$$\begin{aligned}\mathscr{L} &= \mathbb{E}_{\mathbf{x}_n^l,\mathbf{y}_n^l\sim\mathscr{X}^l}\mathscr{L}_S(\mathbf{x}_n^l,\mathbf{y}_n^l)+\\ &\lambda\mathbb{E}_{\mathbf{x}_m^u,\mathbf{y}_m^u\sim\mathscr{X}^u}\{\mathscr{L}_{CSE}(\mathbf{x}_m^u)+\mathscr{L}_{CEC}(\mathbf{x}_m^u)+\mathscr{L}_{PS}(\mathbf{x}_m^u,\mathbf{y}_m^u)\},\end{aligned} \tag{4.13}$$

where λ plays the role in balancing the supervised loss and unsupervised losses. In Eq. (4.13), it is obvious that supervised training becomes dominant with smaller values of λ, and vice versa. Model performance with respect to λ were studied in the experiments section.

In our semi-supervised learning framework, the unsupervised losses were constructed based on self-supervised schemes to minimise the impact of crowd variations, while a supervised loss was designed to learn better crowd information. These proportion of different types of losses should be dynamic at different training stages. Given that, we employed the dynamic λ training strategy, as shown in Algorithm 1.

Algorithm 1: Training Strategy (Dynamic λ)

Input:

- Labeled set $\mathscr{X}^l = \{\mathbf{x}_n^l,\mathbf{y}_n^l\}_{n=1}^{N^l}$;
- Unlabeled set $\mathscr{X}^u = \{\mathbf{x}_m^u\}_{m=1}^{N^u}$;

Output: Model parameters $\boldsymbol{\Theta}$ (corresponding to $\mathscr{F}(\cdot)$, $\mathscr{D}(\cdot)$, and GCRU$(\cdot)$)

Initialisation;
for $b = 1$ *to* B **do**
  Generating pseudo labels: $\{\mathbf{y}_m^u\}_{m=1}^{N^u}$;
  $\lambda = \frac{b-1}{B}$;
  $\mathscr{L}^b = \mathbb{E}_{\mathbf{x}_n^l,\mathbf{y}_n^l\sim\mathscr{X}^l}\mathscr{L}_S(\mathbf{x}_n^l,\mathbf{y}_n^l)+$
   $\lambda\mathbb{E}_{\mathbf{x}_m^u,\mathbf{y}_m^u\sim\mathscr{X}^u}\{\mathscr{L}_{CSE}(\mathbf{x}_m^u)+\mathscr{L}_{CEC}(\mathbf{x}_m^u)+\mathscr{L}_{PS}(\mathbf{x}_m^u,\mathbf{y}_m^u)\}$;
  Calculating gradients: $\nabla_{\boldsymbol{\Theta}}^b \leftarrow \frac{\partial\mathscr{L}^b}{\partial\boldsymbol{\Theta}}$;
  Updating model: $\boldsymbol{\Theta} \leftarrow (\boldsymbol{\Theta}, \nabla_{\boldsymbol{\Theta}}^b)$
end

In Algorithm 1, the joint loss $\mathscr{L}^b$ changes w.r.t training epoch. Specifically, in the first epoch (i.e., $b = 1$), due to the randomly generated pseudo labels, the coefficient of the unsupervised loss is set to zero (with $\lambda = 0$). With the increasing training epoches, the unsupervised loss would become more important (which can reduce the effect of crowd variants), yielding more reliable pseudo labels (for unsupervised learning in further epoches). With a large total epoch, unsupervised loss would nearly have the same weight as the supervised loss, serving as a domain-knowledge driven regularisation term for robust crowd density estimation.

Table 4.1 Semi-supervised algorithm comparison under the settings of Gaussian-process (GP) based method[194]

-	Split(%)	UCF_QNRF		ShanghaiTech A		ShanghaiTech B	
Method	$N^l(\%)$, $N^u(\%)$	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓
GP [194]	5%, 95%	160.0	275.0	102.0	172.0	15.7	27.9
Ours (S4-Crowd)	5%, 95%	**158.6**	**269.6**	**100.8**	**170.3**	**15.3**	**24.4**
GP [194]	25%, 75%	147.0	226.0	91.0	149.0	-	-
Ours (S4-Crowd)	25%, 75%	**140.5**	**224.8**	**81.4**	**137.9**	**12.9**	**21.6**
GP [194]	50%, 50%	136.0	218.0	89.0	148.0	-	-
Ours (S4-Crowd)	50%, 50%	**119.9**	**214.3**	**74.1**	**125.3**	**9.8**	**19.2**
GP [194]	75%, 25%	129.0	210.0	88.0	139.0	-	-
Ours (S4-Crowd)	75%, 25%	**114.6**	**198.7**	**68.3**	**117.3**	**9.7**	**17.8**

4.4 Experiments

4.4.1 Datasets

We evaluated our approach on four popular crowd counting datasets, namely, **ShanghaiTech A/B** [259], **UCF_QNRF** [88], and **World Expo'10 (WE)** [254]. Specifically, **ShanghaiTech** [259] contains two parts, part A and part B. Part A was randomly collected from internet of congested street view with 300 images in the training set and 182 images in the test set. Part B was collected from similar street views with relatively sparse crowd, and it has 400 images in the training set and 316 images in the test set. **UCF_QNRF** [88] is a large-scale dataset with 1535 high-resolution images, with 1201 images in the training set and 334 images in the test set. Both ShanghaiTech and UCF_QNRF were collected in various crowd scenes. **World Expo'10 (WE)** dataset [254] was collected in fixed crowd scenes with 3380 training images and 600 testing images.

For fair algorithm comparison, following other semi-supervised learning approaches in the crowd counting community [134, 194], we kept the test set unchanged while splitting the training set into labeled set (with size N^l) and unlabeled set (with size N^u). More details of the training set splitting can be found in respective experiments below.

4.4.2 Implementation Details

VGG16 with first 13 convolutional layers was a popular backbone for CNN-based crowd counting [33, 93], which was also used in this work (i.e., $\mathscr{F}(\cdot)$). For density map generator $\mathscr{D}(\cdot)$, we used 3 convolutional layers and 1 up-sampling layer. The model was optimised by Adam [99] with batch size 30, and learning rate $1e-5$.

For image transformations (i.e.,$\{\mathscr{A}_k(\cdot)\}_{k=1}^{K}$), we used $K = 5$ image operations including grayscale conversion, gamma adjustment, illumination adjustment (bright/dark), perspective adjustment.

4.4.3 Evaluation Metrics

Mean Absolute Error (MAE) and Mean Square Error (MSE) are two most popular evaluation metrics for performance evaluation [33, 93], which also used in this work. Given the predicted crowd counts of N testing images $\{e_i^{pr}\}_{i=1}^{N}$ (can be inferred from Eq. (4.2) with $\mathscr{F}(\cdot)$ and $\mathscr{D}(\cdot)$) and the corresponding ground truth counts $\{e_i^{gt}\}_{i=1}^{N}$, MAE/MSE can be written as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|e_i^{pr} - e_i^{gt}\right|, MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(e_i^{pr} - e_i^{gt}\right)^2}. \tag{4.14}$$

4.5 Results and Discussions

In this section, we qualitatively and quantitatively evaluate our proposed methods with in-depth discussions.

4.5.1 Comparison with SOTA Methods

To the best of our knowledge, there were three representative works [134, 194, 263] leveraging unlabeled data for improved crowd counting, and two of them were semi-supervised learning[134, 194], which were compared with our method.

Comparison with Gaussian-Process(GP) based method

Sindagi et al. proposed a GP-based semi-supervised method [194] and evaluated it in a few different settings (i.e., in terms of the labeled-set/unlabeled-set split in % of the training set). Following their settings [194], on three datasets UCF_QNRF, ShanghaiTech A, and ShanghaiTech B we compared GP-based approach with our S^4*Crowd* method. As shown in Table 4.1, our method outperformed GP-based method in different settings on the three datasets, and the performance gain tended to be larger on UCF_QNRF and ShanghaiTech A. For ShanghaiTech B, since it is a simple dataset with relatively sparse crowd, the lower error limit can be easily hit. Nevertheless our method still obtained higher results. It is worth to note that with an extremely limited amount of labeled data (i.e., 5%), it is challenging

to train a model as good as in other settings. In order to be fair in real-world scenarios, we randomly select the labeled data in this work and the selection mechanism will be considered in the future, which may be useful for crowd counting in extreme situations.

Table 4.2 Semi-supervised algorithm comparison under the settings of [134]

-	UCF_QNRF		ShanghaiTech A		ShanghaiTech B		WE	
Split	$N^l(721), N^u(480)$		$N^l(90), N^u(210)$		$N^l(120), N^u(280)$		$N^l(947), N^u(2433)$	
Method	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓
L2R [132]	148.9	249.8	90.3	153.5	15.6	24.4	13.9	-
UDA [232]	144.7	255.9	93.8	157.2	15.7	24.1	14.2	-
MT [204]	145.5	250.3	94.5	156.1	15.6	24.5	14.1	-
ICT [212]	144.9	250.0	92.5	156.8	15.4	23.8	14.9	-
IRAST [134]	135.6	233.4	86.9	148.9	14.7	22.9	11.1	-
IRAST(SPN) [134]	128.4	225.3	83.9	140.1	-	-	-	-
Ours (S4-Crowd)	**127.0**	**211.1**	**77.8**	**128.8**	**10.9**	**17.3**	**9.3**	-

Comparison with Inter-Relationship-Aware Self-Training

Liu et al. proposed a semi-supervised crowd counting method [134] by leveraging surrogate tasks, which is Inter-Relationship-Aware Self-Training, and evaluated it on a number of datasets. In their settings, they randomly chose labeled set with size $N^l = 90$ in ShanghaiTech A; $N^l = 120$ in ShanghaiTech B; $N^l = 721$ in UCF_QNRF; and $N^l = 947$ in WE, and used corresponding dataset's rest training set as unlabeled set. These settings as well as a number of semi-supervised learning algorithms' performance can be found in Table 4.2.

Out of the algorithms, Unsupervised Data Augmentation (UDA) [232], Mean teacher (MT) [204] and Interpolation Consistency Training (ICT) [212] were three highly popular semi-supervised methods. For crowd counting tasks, Liu et al. modified them by using the density map as output [134]. Learning to Rank (L2R) [132] was a self-supervised crowd modelling method that can leverage unlabeled crowd images from Internet, and Liu et al. [134] also changed it to the same semi-supervised setting as shown in Table 4.2.

We compared these algorithms as well as IRAST, IRAST(SPN) [134] with our S^4*Crowd* method, and ours outperformed all the algorithms. Compared with the most recent IRAST model [134], our method had a 8.6 MAE reduction on UCF_QNRF; 9.1 MAE reduction on ShanghaiTech A; 3.8 MAE reduction on ShanghaiTech B and 1.8 MAE reduction on WE. It is worth mentioning that Liu et al. also combined their IRAST with state-of-the-art supervised model SPN[29] for further error reduction. In this case, even without external supervised model, our S^4*Crowd* approach still performed better.

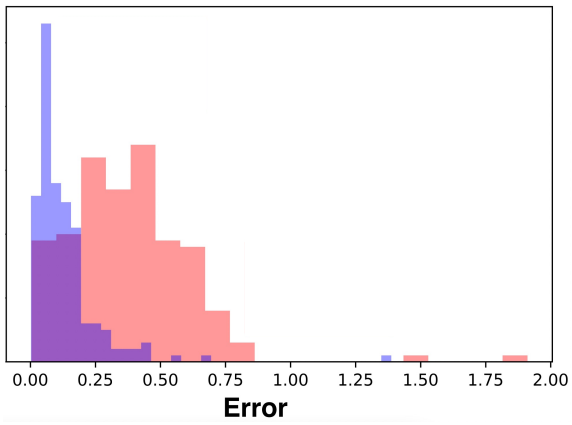


Fig. 4.4 Blue: Errors of VGG16 baseline prediction. Orange: Errors of our pseudo ground truth.

Feasibility of pseudo density

We also conducted a experiment to evaluate the feasibility of using the pseudo ground truth density for unlabeled data. As it is shown in Fig. 4.4, based on ShanghaiTech part A, we visualized the normalized errors of pseudo ground truth density maps and directly predicted density maps, which were compared to the real ground truth. The pseudo density acts as the more accurate supervision of unlabeled data, and using it to train the model can improve the generalibility when labeled data is limited.

4.5.2 Ablation Study

ShanghaiTech A was a popular dataset for ablation studies in the crowd counting research community [6, 93], and it was also used in our study. We empirically set the size (in %) of labeled data (resp. unlabeled data) to $N^l = 15\%$ (resp. $N^u = 85\%$).

Table 4.3 The effect of r for scaling operation $\mathscr{R}(\cdot)$.

ShanghaiTech A			
$\mathscr{R}(\cdot)$	$N^l(\%),\ N^u(\%)$	MAE	MSE
$r = 0.1$	15%, 85%	92.3	159.2
$r = 0.2$	15%, 85%	89.6	159.3
$r = 0.4$	15%, 85%	86.5	155.4
$r = 0.6$	15%, 85%	**82.1**	**143.7**
$r = 0.8$	15%, 85%	88.7	160.2

Effect of re-scaling rate r

The scale variation was modelled with a CSE loss/regularisation in the unsupervised pathway, which was controlled by a re-scaling factor r. Here we focused on the challenging down-sampling scenarios (with low resolution crowd heads/images) by setting the re-scaling factor r at the range of $(0,1)$. Specifically, we studied $r = \{0.1, 0.2, 0.4, 0.6, 0.8\}$, and the results in Table 4.3 suggested that with moderate value ($r = 0.6$), our S^4*Crowd* method reached the best results.

Effect of K ***augmentations***

Our S^4*Crowd* method also modelled different crowd variations by employing K image operations (aka augmentations). Table 4.4 showed the performance with different K image operations. We empirically found that $K = 3$ (by setting images to "grayscale, bright, dark") yielded the lowest MAE.

Table 4.4 The effect of K image operations on modelling crowd variations. K=1 (grayscale), K=2 (grayscale, bright), K=3 (grayscale, bright, dark), K=4 (grayscale, bright, dark, gamma adjustment), K=5 (grayscale, bright, dark, gamma adjustment, perspective adjustment)

ShanghaiTech A			
Method	$N^l(\%),\ N^u(\%)$	MAE	MSE
$K = 1$	15%, 85%	86.1	142.9
$K = 2$	15%, 85%	85.4	**142.1**
$K = 3$	15%, 85%	**82.1**	143.7
$K = 4$	15%, 85%	86.2	151.5
$K = 5$	15%, 85%	95.3	188.5

Effect of each component

Our S^4*Crowd* framework includes many components (such as CSE loss, CEC loss, PS loss, GCRU), and it is important to study the effectiveness of them. In Table 4.5, we can observe supervised approach alone (VGG16) with only limited labels yielded the worst results. On the other hand, VGG16's performance can be substantially improved even with the simplest self-supervised regularisation terms (V+CSE, or V+CEC), which suggested the necessity of leveraging unlabeled data (on modelling crowd variations). Our framework's error can be further reduced with more proposed components (e.g., with PS loss, GCRU).

To study the effect of GCRU, we replaced GCRU by GRU, and reshaped the corresponding high-order-tensor sequences into vector sequences as input. The increased errors suggested the effectiveness of the crowd-driven GCRU in crowd counting tasks. We also changed our backbone from VGG16 to advanced CSRNet [122] and CAN [131], and the lower errors suggested the flexibility of our *S*^{4}*Crowd* framework.

Table 4.5 Effect of each component in our *S*^{4}*Crowd*

ShanghaiTech A			
Method	$N^l(\%),\ N^u(\%)$	MAE	MSE
VGG16(V)	15%, 0%	107.5	211.3
V + CSE	15%, 85%	87.9	155.8
V + CEC	15%, 85%	86.4	158.1
V + GCRU + PS	15%, 85%	88.9	164.9
V + GCRU + PS + CEC	15%, 85%	83.9	150.3
V + GCRU + PS + CSE	15%, 85%	86.4	149.2
V + GRU + PS + CSE + CEC	15%, 85%	84.7	147.0
V + GCRU + PS + CSE + CEC	15%, 85%	82.1	143.7
CSRNet + GCRU + PS + CSE + CEC	15%, 85%	81.8	141.6
CAN + GCRU + PS + CSE + CEC	15%, 85%	**80.3**	**137.4**

Effect of training strategy

We also studied the the effect of different training strategies for our semi-supervised framework *S*^{4}*Crowd*, and the results were reported in Table 4.6. We can see our dynamic λ scheme yielded much lower errors than other strategies which used fixed values of λ.

One explanation is that the importance of unsupervised loss and supervised loss may change during different training stages. At the early training stages, the generated pseudo labels were less reliable and the corresponding importance (measured by λ) should be low. With more training iterations, the model benefited more from the self-supervised CSE/CEC losses (on crowd variations modelling) and supervised loss, and in this case the generated pseudo labels tended to be more reliable. With higher quality pseudo labels, the corresponding PS (Pseudo Supervised) loss should have a higher weight. In our dynamic weight scheme, we can see the weight of the unsupervised loss (PS+CSE+CEC) increased with the more training epoches. At the end of the training, the unsupervised loss would have nearly the same weight as the supervised loss (see Algorithm 1), serving as an important regularisation term in the learning tasks.

Table 4.6 Effect of the training strategy

ShanghaiTech A			
Method	$N^l(\%),\ N^u(\%)$	MAE	MSE
Joint loss in Eq.(4.13) with $\lambda = 0$	15%, 85%	143.4	356.4
Joint loss in Eq.(4.13) with $\lambda = 0.1$	15%, 85%	112.8	240.2
Joint loss in Eq.(4.13) with $\lambda = 0.3$	15%, 85%	105.8	228.9
Joint loss in Eq.(4.13) with $\lambda = 0.7$	15%, 85%	114.8	229.3
Our dynamic λ (Algorithm 1)	15%, 85%	**82.1**	**143.7**

Table 4.7 On leveraging the external unlabeled data

Method	$\mathscr{X}^l(N^l)$	$\mathscr{X}^u(N^u)$	MAE
VGG16	ShanghaiTech A(300)	None(0)	72.8
L2R	ShanghaiTech A(300)	Google(3409)	72.0
Ours (*S*^{4}*Crowd*)	ShanghaiTech A(300)	UCF_QNRF(1201)	**70.3**

Effect of leveraging external unlabeled data

In previous experiments, both $\mathscr{X}^u$ and $\mathscr{X}^l$ were from the same dataset. To simulate the real-world scenarios, we also ran experiments using external unlabeled data. In [132], the authors collected $N^u_{L2R} = 3409$ unlabeled crowd images from Google, and used them as additional information for modelling. For fair comparison, we employed the same network structure as their L2R method [132] (i.e., same $\mathscr{F}(\cdot)$, and a different $\mathscr{D}(\cdot)$ with only 1 convolutional layer). Since the Google crowd images used in [132] were not publicly available, we employed UCF_QNRF as external unlabeled data $\mathscr{X}^u$. We used ShanghaiTech A training set as $\mathscr{X}^l$, and used test set (i.e., with 182 images) for evaluation. The detailed settings and the results were reported in Table 4.7, from which we can see our method yielded lower MAE than L2R, even with a much smaller external $\mathscr{X}^u$. We also noticed both methods did not outperform the supervised baseline substantially. One possible reason can be the differences from two different data sources, and domain adaptation will be explored to boost the performance in the future.

4.6 Conclusion

In this paper, we proposed a semi-supervised framework *S*^{4}*Crowd* for crowd counting tasks. Two self-supervised regularisation terms CSE/CEC, pseudo labeling, GCRU, were proposed as key components in this framework together with a dynamic training scheme for better crowd modelling. We comprehensively studied the effectiveness of each component, and

our method outperformed other state-of-the-arts in various semi-supervised settings on public datasets. In the future, we will explore domain adaptation methods on leveraging the unlabeled crowd data in the wild.

Chapter 5

Unified Attention Model for Visual Feature Modelling

In this chapter, we explore the advancement of deep learning in computer vision, focusing on an innovative integration of attention mechanisms within neural networks. Our model, inspired by both supervised and unsupervised learning approaches, merges the strengths of Convolutional Neural Networks and Transformer-based Networks in a parallel design. This approach enhances the model's ability to process and prioritize crucial visual information at both local and mid-level scales along with global-level information. Key to our design is the Dynamic Local Enhancement module, which selectively emphasizes informative visual patches, and the Unary Co-occurrence Excitation module, which identifies and maintains important spatial relationships between patches. Together, these modules enable our model to effectively bridge the local-mid-global processing gap and adapt to a variety of complex vision tasks, setting a new standard in deep vision recognition.

5.1 Introduction

Supervised learning with annotated data and unsupervised learning with unlabelled data is the main paradigm in deep learning, in which our proposed attention mechanism is effective, we further desire to investigate and propose a novel attention mechanism that can be beneficial to varied vision recognition tasks. Hence, this chapter aim to propose a unified deep vision recognition model based on the attention mechanism. Recent deep model design has become the essential computer vision task [59]. Embracing the power of high-performance computing and rich visual contents from online platforms, the current leading paradigm of computer vision aims to pre-train a large-scale, multi-task, multi-modality

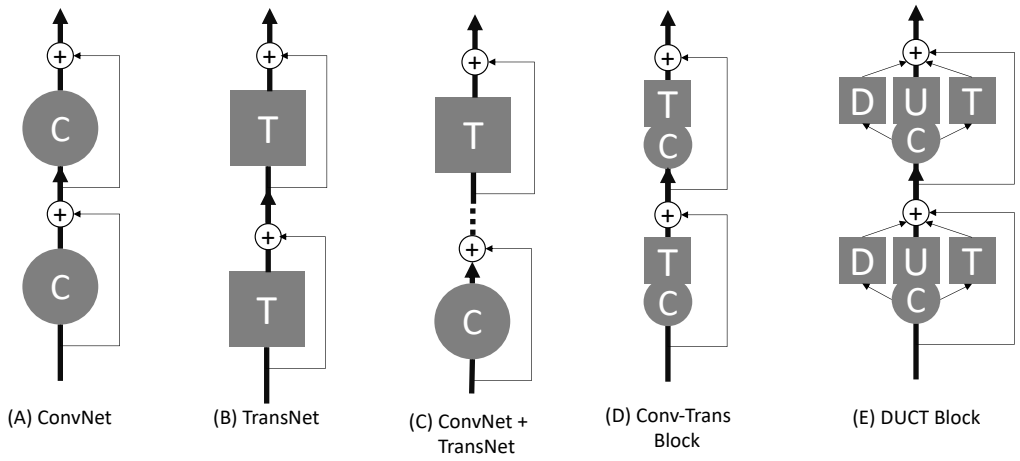


Fig. 5.1 Figure Comparison of existing convolution network [84] (A) and transformer [210](B) architecture designs with the proposed DUCT blocks, which consists of Dynamic local enhancement module, Unary co-occurrence excitation module, conventional Transformer layer(multi-head self-attention) and Convolution. While previous work integrates convolution and transformer layers in a separate series [243] (C), recent trends alternate transformer and convolution in a block-wise way [228](D). Our DUCT (E) is the proposed parallel structure combining a dynamic local enhancement module, a unary co-occurrence excitation module, and multi-head self-attention in a block-wise design.

model that can be transferred to down-stream tasks. While Convolution-based deep neural Network (ConvNet) architectures have established a leading position in key computer vision tasks, e.g., image detection, classification, segmentation, the community has been seeking for multi-modal solutions since the last decade. An inevitable and essential topic is about sequential-modeling which has natural applications in videos, free-texts, audios, and many other signals of wearable devices. The traditional RNN-based paradigm was challenged by the transformer-based neural Network (TransNet) architecture [210] which has soon become a dominant approach. Recent research has shown that the TransNet architecture can even outperform ConvNet on pure vision tasks [51]. Debate has focused on whether the vision and language tasks should be brought together, and the model paradigm should be unified in the new TransNet formula for better transition between multi-modal tasks.

ConvNets have natural advantages in visual tasks due to their spatial prior. The existing TransNet paradigm breaks visual data into local patch tokens. The natural 2D or 3D neighborhood dependence is broken into a 1D sequential order. Fully-connected attention with

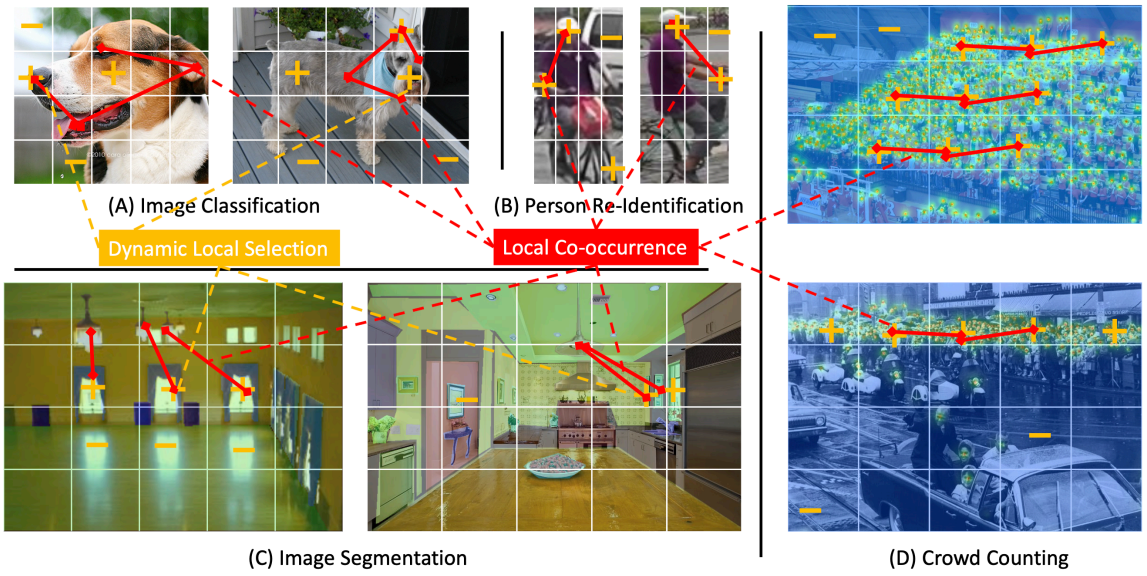


Fig. 5.2 Illustration of our proposed Dynamic Local Enhancement (DLE) and Unary Co-occurrence Excitation (UCE) in different computer vision tasks. DLE aims to assign weights to important local patches for convolution (in orange colour). UCE searches for unique co-occurrence between a local patch and others. Such co-occurrence at the feature-map level can achieve higher invariance. DLE, UCE and multi-head self-attention are combined to detect local, mid-level and global information in a complementary way.

dense tokens is needed to capture the dependence, which makes TransNets suffer from poor scalability and flexibility. A few recent attempts introduce convolution to transformers and achieved promising results [228]. As it is shown in Fig. 5.1, most exiting deep architectures adopt residual connections. Layers between two residual connections can be regarded as a block. For simplicity, the figure does not include normalization, activation, pooling, etc. Most TransNet also adapts residuals as shown in Fig. 5.1 (B). A straightforward approach is to break the low-level vision tasks using ConvNets and apply TransNet onto the feature maps to process the high-level information. Dai et al. has proposed a CoatNet that consists of consecutive conv blocks followed by further transformer layers [43]. This paradigm is illustrated in 5.1 (C). Further attempts concentrate on introducing convolution to transformers as a hybrid block as shown in 5.1 (D). It is intuitive to see that existing attempts for Conv-TransNet follow the design in a series order. The ConvNet before the transformer layer is interpreted as a tokenizer feature extraction of each image's local patch. The resulting feature maps are composed of the global information. Without further constraint or prior information, the following attention layer needs to compute the dependency between all of these feature maps. This property is particularly susceptible when the global structure

of local patches is severely shifted, e.g. incomplete or occluded objects, or due to unseen viewing angles on the test domain.

In this chapter, we explore parallel design to enhance the local and mid-level information as complementary to the global attention model, where the fusion of diverse features is hypothesized to boost the model performance on various scenarios. Our motivation is illustrated in Fig. 5.2. In most computer vision tasks, a few local patches are much more informative than others. Also, some patches can cause ambiguity and should be suppressed. When computing the global attention in an unnatural sequential order, the response from such local patch information is negated by other tokens, resulting in a blurring effect. Therefore, we first introduce a *Dynamic Local Enhancement* (DLE) module to achieve dynamic local selection, i.e., force the model to assign higher weights to important patches. For example, as illustrated in Fig. 5.2 (C), the reflection of window light on the floor can confuse the model and thus needs to be assigned lower weights. In contrast, the actual window region will receive high weights so that the convolution signal can be safely preserved, complementing the global attention.

The other novel idea comes from our observation of the local co-occurrence property. For example, the local patch of dog eyes often occurs within the (or nearby/alongside the) nose and mouth area. Such a correlation is very sensitive to shifting, e.g., view angles, occlusion, etc. Similarly, the crowd counting problem—shown in Fig. 5.2 (D)—is where crowded patches with similar heads are highly associated, and also associated with other crowds compared to non-crowded ones.

Taking advantage of the unnatural order of patch tokens, we can compute the affinity matrix of the patch tokens. Each row of the affinity matrix then represents the 1-to-n correlation for each of the local patch tokens. We develop a novel *Unary Co-occurrence Excitation* (UCE) module on the 1-to-n correlation vector. The key idea is that relative correlation can hold regardless of whether the positions are changed. As it is shown in Fig. 5.2 (C), the windows often co-occur with the lamp. In the two compared images, the patch locations of windows and lamps are very different. But the pair-wise or group-wise correlation score can hold for the tokens of windows and lamps regardless of the position shifting. It is named 'unary' because each patch token is only assigned with a single unique convolutional kernel, to search for similar score patterns. Also, it aims to search for invariant groups or pair-wise token correlations for each local patch. Such groups consist of several correlated patch tokens as a part of the global structure. Therefore, Unary Co-occurrence Excitation can provide mid-level information as a bridge between the local and global gaps. We summarize our main contributions as follows:

- The first attempt to integrate parallel structure within a hybrid Conv-Trans block.

- We introduce a dynamic local enhancement module to preserve highly informative local patch/token information.
- We propose a novel unary co-occurrence excitation module that searches for position-invariant local co-occurrence, achieved by convolution over group-wise correlation scores between patch tokens.
- The dynamic unary enhancement with Transformers is combined as a 3-channel block. And an adaptive patch merging process is designed to select diverse features and reduce redundancy. Finally, the DUCT deep architecture (consisting of aggregated DUCT blocks) is comprehensively evaluated in four essential computer vision tasks, i.e., image-based classification, segmentation, retrieval, and regression (density estimation). The proposed method outperforms existing Conv-Trans design in series with state-of-the-art results.

The proposed DUCT block and the parallel design aims to bring new theoretical insights and help future work build a large-scale architecture for extremely large datasets. Yet the evaluation on lots of large-scale datasets is out of the scope of this chapter; our goal is to design a flexible and generic Conv-TransNet on different computer vision paradigms. The following chapter is organized as follows. In section two, our literature review examines both pure transformers in vision and hybrid visual transformers. Our technical details and methodology are introduced in section three. In section four, we introduce the experimental design and results discussion of the model performance on four computer vision tasks. Theoretical statements are supported by both a qualitative and quantitative ablation study. Our work is summarized in the last section, where further work and potential impacts are discussed.

5.2 Related Work

Transformers benefit from the multi-head self-attention mechanism, and have become the prominent model in natural language processing (NLP) [173], allowing for information capture over different ranges. Recently, transformers and their variants have shown encouraging potential on computer vision tasks, and are considered to be alternative models to classical convolutional neural networks (CNNs). Here, we aim to summarize and discuss the recent development of pure transformer models and hybrid transformers.

5.2.1 Transformers in Computer Vision

Earlier research largely focuses on the differences between words and pixels, with various methods that apply the word embedding concept to image data. The interdependence among pixels is critical to be captured by the self-attention mechanism, hence Parmar et al. [163] conducted experiments for image generation tasks by applying self-attention for each query pixel instead of modeling them globally. With the pixel-wise channel-specific embedding and multi-head self-attention blocks, the proposed model achieved competitive performance on image recognition task without using extra convolutional blocks.

An alternative method to scale attention is to apply it in blocks of varying sizes; Cordonnier et al. [39] applied self-attention on top of 2x2 patches, although this does not generalize to large-scale vision tasks. Also, some works tried to increase the receptive field with different sizes of patches to flexibly handle larger resolution images. In recent years, transformers have emerged as the backbone that drives advances in image classification—traditionally dominated by CNNs. Dosowitzky et al. [51] proposed a Visual Transformer (ViT), which has a similar form to those used in NLP tasks. It performs well on image classification tasks directly applied to image patch sequences. The network adopts a similar approach to BERT's tokenization method, where a learnable embedding is applied to the sequence of embedding patches. The state of this embedding serves as the image representation. In addition, a learnable 1D positional embedding was added to the patch embedding to retain positional information. In most cases, ViT is pre-trained on large datasets such as the ImageNet and then fine-tuned for smaller downstream tasks. Beyond the ViT, a set of variants were proposed to improve the performance; mainly focusing on enhancing locality, improving self-attention performance and architecture design. For excavating local information in different scales and locations, TNT [70] further divides the patch used in ViT into multiple sub-patches, where a transformer-in-transformer architecture was developed to capture the relationship between such inner transformer blocks, and for patch-level information exchange an outer transformer block was developed. Swin Transformers [136, 50] conduct local attention within a window and introduce a shifted window partitioning approach for cross-window connections. Shuffle Transformer [85] further utilizes the spatial shuffle operation instead of shifted window partitioning to achieve cross-window connections. RegionViT [23] generates regional tokens and local tokens from an image, and each forward token receives global information via attention with regional tokens. However, for vision tasks, recent work suggests that the transformers focus more on the global features, where it is still uncertain if other lower levels of information are necessary. In this work, we observed that different levels of information are still critical in a vision recognition network. This motivated us to design a

new paradigm of hybrid transformer network to enhance representation learning inside and across the different tokens.

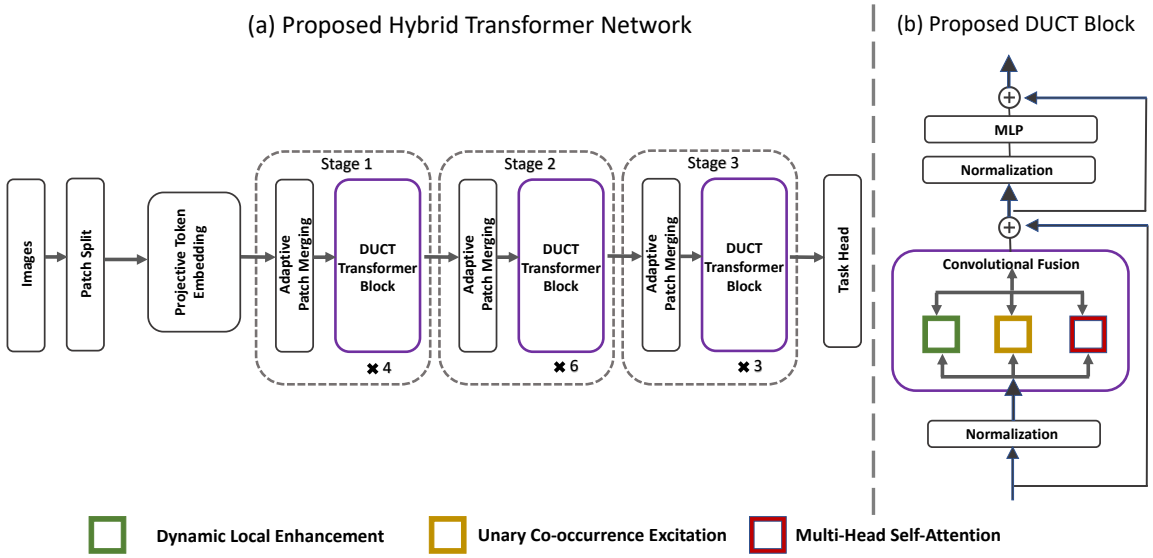


Fig. 5.3 (a) Architecture overview of the proposed hybrid transformer network DUCT. (b) The proposed hybrid transformer block for DUCT.

5.2.2 Hybrid Vision Transformers

Many recent works have independently confirmed that transformers can be successfully applied to various vision tasks [97, 1, 110, 200, 197, 248, 184], as they are able to capture long-range dependencies in inputs. However, there are still gaps in performance when compared with traditional CNN-based networks. The performance of ViT is largely limited where the training data is inadequate, especially compared with that of state-of-the-art ConvNets. There have been some works that combine convolution with self-attention in recent years. Although it seems that tokenized embedding works well, the transformers still need to be enhanced to learn dense, repeatable patterns (e.g., textures and edges) which convolutions are significantly more efficient at learning. And these early frameworks generally require a significant increase in computing resources to outperform convolutional variants.

There has also been recent interest in combining CNNs with forms of self-attention. Existing research focuses on improving the capability of extracting local information. With the image domain-specific inductive biases, [228] proposed the CvT to combine CNNs and transformers to model both local and global dependencies for image classification in an efficient way. Their model consists of two novel structural changes. Firstly, they use

convolutional projection modules to replace the existing position-wise linear projection for the attention operation. Secondly, they adopt a hierarchical multi-stage structure to support varied resolutions of 2D reshaped token maps, named convolutional token embeddings. Other works similarly analyzed the drawbacks of directly applying transformers from NLP on image tasks, such as [243, 120, 67]—focusing on either replacing or combining the feedforward network (FFN) with convolutional layers in each transformer module to better capture the correlation between neighboring tokens.

It's worth mentioning that there is also research into leveraging self-attention-style techniques to boost the performance of CNNs; [11] augment convolutions by concatenating convolutional feature maps with explicit self-attention. This additionally validates the benefits of combining both architectures. To enhance the model's awareness of global information, Wang et al., [223] proposed non-local operations as a family of building blocks that can capture long-range dependencies from sequences. Their approach achieves more accurate classification results for videos than 2D and 3D ConvNets, and is efficient in utilizing computational resources. The next year [81] introduced a geometric prior on the new local relation layer; the self-attention based layer extracts more representative compositional structures and adapts aggregation weights according to the spatial context. Most of the existing hybrid transformer paradigms attempted learning via series stream information, whereas investigating the parallel order of integrating information is an area that has not yet been well explored. This work aims to design a novel parallel hybrid transformer paradigm, where the goal is to enhance local, mid and high-level information; we hypothesize such diverse features are able to boost model performance in various vision tasks.

5.3 Methodology

The proposed approach aims to assemble off-the-shelf mainstream deep learning components in the most appropriate way to accomplish their mutual complementarity. Specifically, the details of each component in the proposed DUCT network will be outlined in the following sections; the Dynamic Local Enhancement module (DLE), the Unary Co-occurrence Excitation (UCE) module, the Multi-Head Vision Transformer (MHVT), together with discussions on the convolution operations accordingly.

5.3.1 Projection-enhanced Transformer

Vision Transformers [96] introduce a way to process input images in raster order, akin to word embeddings in transformers for NLP tasks. They use self-attention to substitute

convolutional operations. Formally, given an input RGB image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ (H and W represent the height and width of the image respectively), the image is partitioned to generate N square patches (*a.k.a.*, tokens) where each patch has a spatial resolution of $\sqrt{P} \times \sqrt{P}$ and $N = \frac{H \times W}{\sqrt{P} \times \sqrt{P}}$. Note that there is no overlap between the adjacent patches. The resulting patches are then flattened and stacked to form $\mathbf{X} \in \mathbb{R}^{N \times P \cdot C}$, where C is the channel of each patch. According to the original Vision Transformer [96], $\mathbf{X}$ is linearly projected to a new embedding space of dimension $N \times C'$ to learn the global dependencies of tokens. However, the linear projection might potentially overlook some useful information because it can only reflect the local patterns represented by the split patches within their limited context.

Projective Token Enhancement To alleviate the impact of linear token embedding, a projective token enhancement module is proposed; a given input RGB image is split into non-overlapping patches $\mathbf{I'} \in \mathbb{R}^{(\frac{H}{4} \times \frac{W}{4}) \times C'}$, which implies that the dimension of the flattened token is $C' = 4 \times 4 \times 3$. The resulting patches are then fed into a projective token embedding module, similar to Swin Transformer [136]—composed of three linear mapping layers and normalization layers—but introduces an additional non-linear activation layer and the standard residual connection to generate the preliminary features of tokens, denoted as $\mathbf{X} \in \mathbb{R}^{N \times D}$.

These features, from the embedded tokens, allow for the low-level cues in each token to be preserved within the hierarchically designed embedding module, complimenting the convolution architecture. This early feature processing is distinct from previous work [51, 243], in that we use the transformer block to directly perform feature extraction on the embedded tokens.

Self-Attention Mechanism The use of self-attention [96, 210] allows for capturing global contextual dependencies present in the N entries of embedded features $\mathbf{X} \in \mathbb{R}^{N \times D}$. Specifically, $\mathbf{X}$ is encoded as the query $\mathbf{Q}$, the key $\mathbf{K}$, and the value matrix $\mathbf{V}$ of dimensions D_q, D_k, and D_v respectively. These act as the input of the self-attention layer. The output of the self-attention layer is a weighted sum of the values:

$$\text{Attention} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{D_k}}\right)\mathbf{V}. \tag{5.1}$$

In other words, the weight for each value is assigned by the scaled dot-product of each query and all keys.

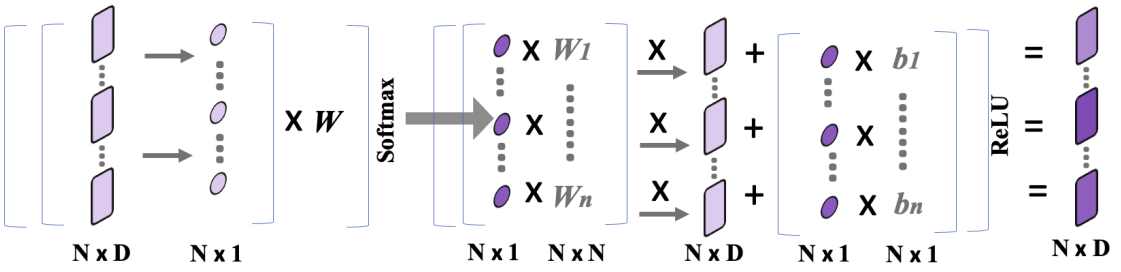


Fig. 5.4 The proposed Dynamic Local Enhancement (DLE) module. Given the token features, it first summarizes the average response, which is transformed to be the attention score. Then the attention score is used to calculate the dynamic convolution kernel for the dynamic local enhancement function (The N denotes the number of the square patches and the D denotes the dimension of the embedding features.).

Multi-Head Self-Attention Self-attention can be decomposed into multiple heads to support parallel and independent computation while considering the diversity of contextual information between patches and the aggregation of different representation subspaces. Specifically, with h as the number of attention heads and the learnable projection matrices $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, $\mathbf{W}_i^V$ and $\mathbf{W}^O$, Multi-Head Self-Attention (MHSA) is calculated in parallel:

$$\begin{aligned} \mathrm{MHSA}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right) &= \mathrm{Concat}\left(\mathrm{head}_1,\ldots,\mathrm{head}_h\right)\mathbf{W}^O \\ \text{where } \mathrm{head}_i &= \mathrm{Attention}\left(\mathbf{Q}\mathbf{W}_i^Q,\mathbf{K}\mathbf{W}_i^K,\mathbf{V}\mathbf{W}_i^V\right), \end{aligned} \tag{5.2}$$

and the output from the multi-head self-attention module is denoted $\mathbf{X}^{\tilde{\mathrm{T}}} \in \mathbb{R}^{N\times D_v}$.

MHSA enables the Vision Transformer to capture global dependencies of the generated tokens without recursion. However, this comes at the expense of scalability in various computer vision tasks, which often require the proposed model to have a more targeted response to the task [174]. For example, as stated in [106], the transformer can only acquire effective local information by training on large-scale datasets (even larger than ImageNet). To this end, as shown in Fig. 6.2, we introduce a dynamic local enhancement and unary co-occurrence excitation module induced from standard convolutional operators to further enhance the expressiveness of features at different scales, thereby improving the modelling capability of the transformer for various tasks.

5.3.2 Dynamic Local Enhancement

Convolution-based deep learning models can extract local pixel information by means of using small filters, while the transformer blocks cannot explicitly model such fine-scale in a

way that is scalable [174, 136, 23]. To enhance the ability of extracting the local features in each patch, a Dynamic Local Enhancement (DLE) module is presented to adaptively estimate a set of learnable convolution kernels that can independently model the relevant spatial information for an individual token (shown in Fig. 5.4). Given the availability of the generated token features $\mathbf{X} \in \mathbb{R}^{N \times D}$, the statistical information represented for each token can be reduced by averaging each row vector of $\mathbf{X}$:

$$\mathbf{x}_n^S = \frac{1}{D} \sum_{d=1}^{D} \mathbf{x}_n (1, d), \tag{5.3}$$

where $\mathbf{x}_n^S$ denotes the output of n_{th} row vector of $\mathbf{X}$. The overall representation $\mathbf{X}^S \in \mathbb{R}^{N \times 1}$ can be formed by stacking the averaged outputs of N row vectors, which is stable to the variations exhibited in the original features $\mathbf{X}$.

The values in $\mathbf{X}$ that summarize the average responses for all tokens are transformed to a set of attention scores by:

$$\mathbf{G}^S = \text{Softmax}\left(\mathbf{W}_2 \left(\tau \left(\mathbf{W}_1 \mathbf{X}^S\right)\right)\right), \tag{5.4}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are two weighting parameters, and τ is the ReLU activation function. Based on obtained attention score $\mathbf{G}^S \in \mathbb{R}^{N \times 1}$, we dynamically estimate the learnable kernels in order to enhance the variability of such attentive responses, where:

$$\mathbf{G}^{\mathbf{w}} = \sum_{n=1}^{N} \mathbf{G}_n^s \tilde{\mathbf{W}}_n, \quad \mathbf{G}^{\mathbf{b}} = \sum_{n=1}^{N} \mathbf{G}_n^s \tilde{\mathbf{b}}_n, \tag{5.5}$$

and $\tilde{\mathbf{W}} \in \mathbb{R}^{N \times N}$, $\tilde{\mathbf{b}} \in \mathbb{R}^{N \times 1}$ are the learnable weights and biases respectively. These are used to dynamically estimate the convolution kernel for different tokens. Furthermore, $\sum_{k=1}^{K} \mathbf{G}_k^s = 1$ with $0 \leq \mathbf{G}_k^s \leq 1$.

Based on the aggregation matrices $\mathbf{G}^{\mathbf{w}}$ and $\mathbf{G}^{\mathbf{b}}$, the locally enhanced features are obtained by:

$$\mathbf{X}^{\tilde{\mathrm{D}}} = \tau \left(\mathbf{G}^{\mathbf{w}} \mathbf{X} + \mathbf{G}^{\mathbf{b}}\right), \tag{5.6}$$

where τ is the ReLU activation function. $\mathbf{G}^{\mathbf{w}}$ and $\mathbf{G}^{\mathbf{b}}$ denote the matrix transformation of a 1D convolution. In contrast to traditional 1D convolution, the weights of the convolution are dynamically assigned by $\mathbf{G}^{\mathbf{w}}$ and $\mathbf{G}^{\mathbf{b}}$. The resulting features $\mathbf{X}^{\tilde{\mathrm{D}}}$ are of shape $N \times D_d$. Under the premise of retaining the feature size of the input token, the proposed DLE module can greatly increase the sensitivity of the local informative features and expand the network's ability to capture diverse information.

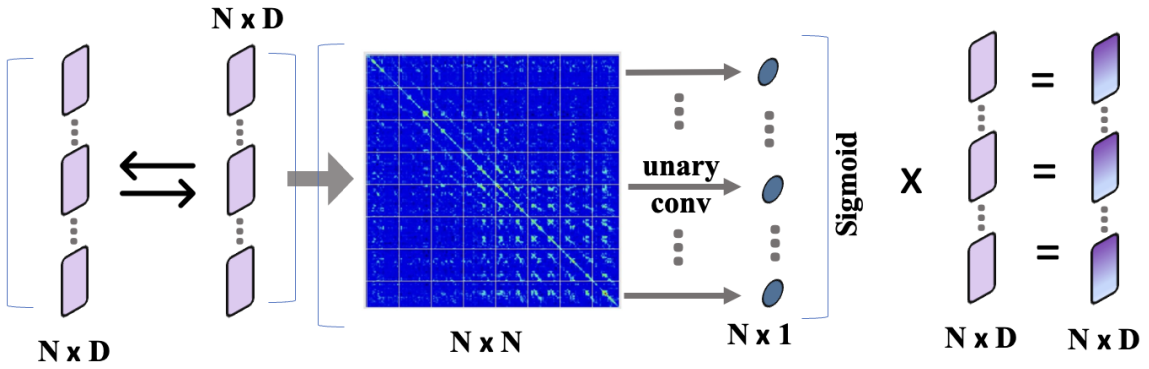


Fig. 5.5 The proposed Unary Co-occurrence Excitation (UCE) module. A correlation matrix is first calculated, and then it is transferred to the attention matrix by a unary convolution, which is used to enhance the 1-to-*n* correlation.

5.3.3 Unary Co-occurrence Excitation

Following the aforementioned observation of local co-occurrence, it is crucial for the model to learn the 1-to-*n* patterns to ensure that the correlations for different combinations of tokens are diverse regardless of changes in patch positions. To achieve this goal, we propose a novel Unary Co-occurrence Excitation (UCE) module, shown in Fig. 5.5. Considering the embedded features of tokens as $\mathbf{X} \in \mathbb{R}^{N \times D}$, the correlations between the tokens and channels are calculated as:

$$\mathbf{M} = \mathbf{X}\bar{\mathbf{I}}\mathbf{X}^{\top}, \tag{5.7}$$

where $\bar{\mathbf{I}} = \frac{1}{D}\left(\mathbf{I} - \frac{1}{D}\mathbf{1}\right)$ with an identity matrix $\mathbf{I} \in \mathbb{R}^{D \times D}$ and a matrix of ones $\mathbf{1} \in \mathbb{R}^{D \times D}$. The obtained correlation matrix $\mathbf{M} \in \mathbb{R}^{N \times N}$ can better reflect the pair-wise 1-to-*n* relationships among all tokens. More specifically, the diagonal entries of $\mathbf{M}$ represent the variances and the other entries represent the covariances between the tokens, meaning that the dependencies of the corresponding token with all other tokens are incorporated. Then, a unary convolution is proposed to effectively encode such 1-to-*n* relationships. Specifically, the matrix $\mathbf{M}$ is reshaped into $\mathbf{M} \in \mathbb{R}^{1 \times N \times N}$, enabling convolution kernels $\mathbf{K}$ of size $1 \times N$ to be applied. This can be expressed by:

$$\tilde{\mathbf{M}} = \sigma\left(\mathbf{K}_{1,N,N} \cdot \mathbf{M}_{1,N,N}\right), \tag{5.8}$$

where σ is the sigmoid function, and $\cdot$ denotes the dot product indicating the convolution operation. The output of the UCE module $\mathbf{X}^{\tilde{\mathrm{U}}}$ is then just a product between the reshaped $\tilde{\mathbf{M}} \in \mathbb{R}^{N \times 1}$ and $\mathbf{X} \in \mathbb{R}^{N \times D}$:

$$\mathbf{X}^{\tilde{\mathrm{U}}} = \mathbf{X}\tilde{\mathbf{M}}. \tag{5.9}$$

Consequently, the resulting feature representations depicted above are concatenated to form a unified representation because they have identical dimensions. Formally, it is represented as:

$$\hat{\mathbf{X}} = \text{Conv}\left(\text{Concat}\left(\mathbf{X}^{\tilde{\text{D}}}, \mathbf{X}^{\tilde{\text{U}}}, \mathbf{X}^{\tilde{\text{T}}}\right)\right), \tag{5.10}$$

where $\mathbf{X}^{\tilde{\text{T}}}$ is the output from Multi-Head Self-Attention (MHSA). The concatenated representation has the shape $3N \times D$, which contains extensive useful information at different levels (low, mid and high levels), and the convolution operation is applied to filter the most valuable information while dropping the redundant information; the final output is then obtained, denoted as $\hat{\mathbf{X}} \in \mathbb{R}^{N \times D'}$.

5.3.4 Adaptive Patch Merging

Inspired by the recent work [136, 228, 222], a patch merging module is applied to combine the distinct feature representations. However, as reported in [140], merely applying regular grid-aware convolutional operations on the reshaped token sets [228, 222] may completely neglect the fact that different tokens usually contribute unequally, and also that tokens may have differing levels of interaction between each other. To handle these issues, motivated by deformable convolution [41], a group of offsets is introduced to effectively sample those informative tokens adaptively and then influence the process of merging tokens.

More formally, the unified features $\hat{\mathbf{X}} \in \mathbb{R}^{N \times D'}$ generated from Eq. (5.10) can be reshaped into $\hat{\mathbf{X}} \in \mathbb{R}^{1 \times N \times D'}$. The standard convolution operation at the location $\mathbf{k}$ of each pixel can then be expressed as:

$$\bar{\mathbf{X}}(\mathbf{k}) = \sum_{\mathbf{k}_i \in [K \times K]} \mathbf{W}(\mathbf{k}_i) \cdot \hat{\mathbf{X}}(\mathbf{k} + \mathbf{k}_i), \tag{5.11}$$

where $\mathbf{k}_i$ enumerates the sampling locations in a convolution kernel (with size $K \times K$). A learnable offset $\Delta\mathbf{k}_i$ is then introduced into Eq.(5.11), yielding the adaptive patch merging scheme:

$$\bar{\mathbf{X}}(\mathbf{k}) = \sum_{\mathbf{k}_i \in [K \times K]} \mathbf{W}(\mathbf{k}_i) \cdot \hat{\mathbf{X}}(\mathbf{k} + \mathbf{k}_i + \Delta\mathbf{k}_i), \tag{5.12}$$

where the learnable offset $\Delta\mathbf{k}_i$ is estimated by an extra independent convolution layer. The output features are additionally transformed via a convolutional layer, a batch normalization layer, and then a GELU activation function. The dimensionality N is decreased to $\frac{1}{4}N$, and D' is increased accordingly to provide more channels/features information as in traditional convolutional neural networks. As the network depth increases, patch merging is used to reduce the number of tokens and control the channel dimension [51, 136] via adaptively

Table 5.1 Comparisons with state-of-the-art methods on ImageNet-1K [106]

Method Type	Network	#Param.(M)	Image Size	FLOPs (G)	top-1 (%)
Convolution Neural Networks	ResNet-50 [74]	25	224 × 224	4.1	76.2
	ResNet-101 [74]	45	224 × 224	7.9	77.4
	ResNet-152 [74]	60	224 × 224	11	78.3
	RegNetY [172]	39	224 × 224	8	81.7
	EfficientNet [202]	19	380 × 380	4.2	82.9
Transformers	ViT-B/16 [51]	86	384 × 384	55.5	77.9
	ViT-L/16 [51]	307	384 × 384	191.1	76.5
	DeiT-S [207]	22	224 × 224	4.6	79.8
	DeiT-B [207]	86	224 × 224	17.6	81.8
	PVT-Small [222]	25	224 × 224	3.8	79.8
	PVT-Medium [222]	44	224 × 224	6.7	81.2
	T2T-ViTt-14 [244]	22	224 × 224	6.1	80.7
	T2T-ViTt-19 [244]	39	224 × 224	9.8	81.4
	TNT-S [70]	24	224 × 224	5.2	81.3
	TNT-B [70]	66	224 × 224	14.1	82.8
	Swin-T [136]	28	224 × 224	4.5	81.3
	Swin-S [136]	50	224 × 224	8.7	83.0
Convolution + Transformers	CvT-13 [228]	20	224 × 224	4.5	81.6
	CvT-21 [228]	32	224 × 224	7.1	82.5
	CoAtNet [43]	25	224 × 224	–	81.6
	MobileViT [144]	5.6	256 × 256	–	78.4
Ours (Hybrid Transformer)	DUCT_{224}	31	224 × 224	12	**83.1**
	DUCT_{384}	31	384 × 384	43.1	**84.7**

integrating the informative patches, which allows for robust hierarchical representations giving the final output.

5.4 Experiments

Our experiments are conducted on four principal computer vision tasks including classification, segmentation, retrieval (person re-identification) and regression (crowd counting).

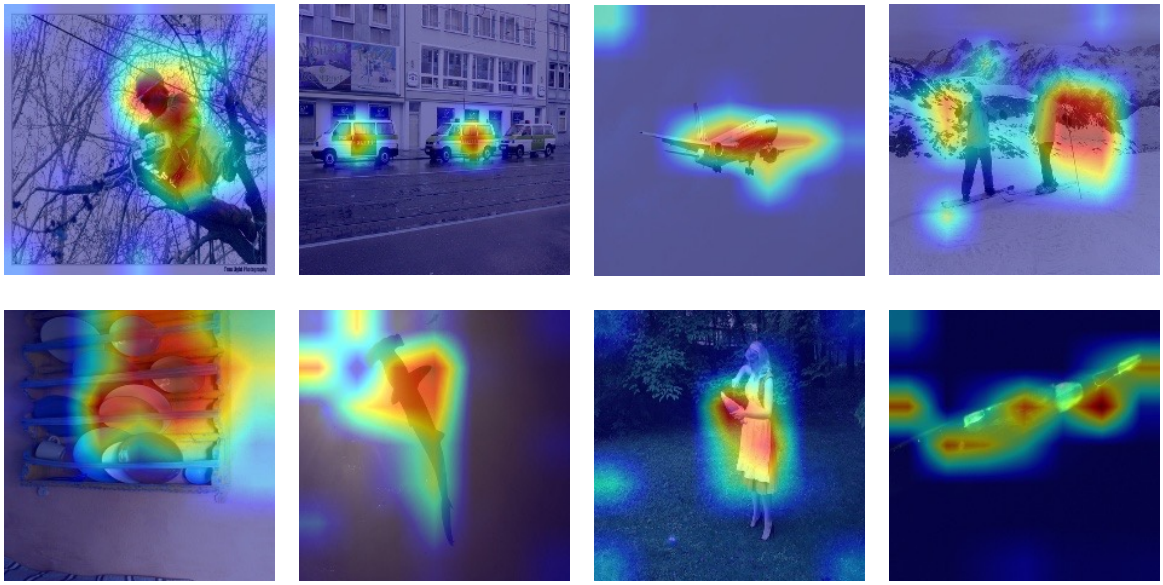


Fig. 5.6 Examples of class response maps from the output to the input on the ImageNet1K dataset.

5.4.1 Model Configurations

Our model receives images of size 224×224 as input, which are initially partitioned into 4×4 patches. Then three linear embedding layers with normalization and residual connections are employed to preserve the subtle local information, and the output sequential token features are input to 3 hybrid transformer stages using the proposed DUCT blocks. The details of the DUCT blocks are described below:

- Stage 1: Patch size is 2 and the channel dimension is 128, the number of MHSA heads is 4, the number of transformer blocks is 4.
- Stage 2: Patch size is 2 and the channel dimension is 320, the number of MHSA heads is 6, the number of transformer blocks is 6.
- Stage 3: Patch size is 2 and the channel dimension is 512, the number of MHSA heads is 8, the number of transformer blocks is 3.

5.4.2 Image Classification

The proposed DUCT is evaluated on five classification benchmark datasets, which are ImageNet-1K [106], CIFAR-10 [105], CIFAR-100 [105], Oxford Pet [162] and Oxford Flowers [155]. These experiments are set up as follows:

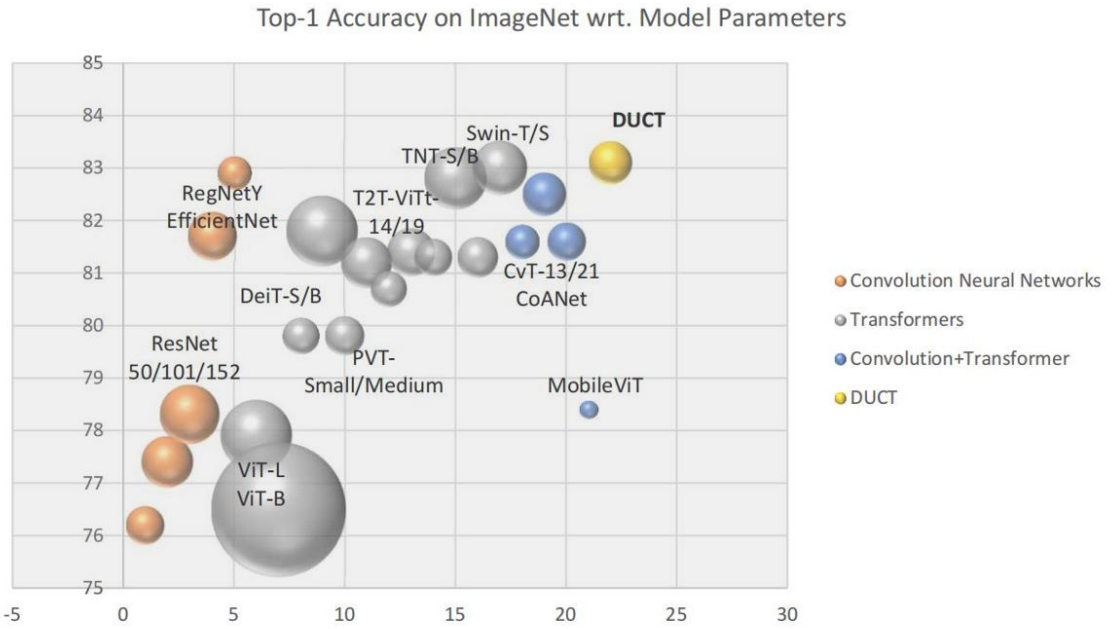


Fig. 5.7 The top-1 accuracy on ImageNet-1K [106] compared to other methods with respect to model parameters.

- ImageNet-1K [106] is a large-scale dataset which contains 1.28M training images and 50K validation images from 1,000 classes. The AdamW optimizer [138] with the cosine decay learning rate scheduler is used to optimize the network. The model is trained for 300 epochs with a batch size of 1024. The initial learning rate for training the entire model is set to 1e-3, and the weight decay is 0.005. The learning rate for adaptive patch merging is separately set to 1e-5. The top-1 accuracy and the computational costs are summarized for comparison.

- CIFAR-10 [105] has 50,000 training images and 10,000 testing images. CIFAR-100 [105] has 100 categories, each with 500 training images and 100 testing images per class. Oxford-Pets [162] has 37 categories, each with about 200 images per class of which 50 are for training, 50 for validation, and 100 for testing. Oxford-Flower [155] contains 102 flower categories and each class has around 40-258 images. We follow the previous work [155, 228] to split the train/validation/test sets. The model is fine-tuned on the model that was pretrained on ImageNet1K. The backbone is optimized using the SGD optimizer with a learning rate of 1e-4 and momentum of 0.9. It is trained for 200 epochs with input size 224×224 and batch size 256.

Table 5.1 shows the performance of DUCT on ImageNet1k compared with existing state-of-the-art methods based on CNNs, transformers and convolutional transformers. Also, a concise overview is shown in Fig. 5.7, based on the top-1 accuracy with respect to model parameters.

Compared with most state-of-the-art convolutional neural networks, the proposed DUCT achieves significantly higher top-1 accuracy. It also can be seen that DUCT achieves a better trade-off between accuracy and speed than existing CNN-based models. EfficientNet [202] and RegNetY [172] are the most recent mainstream convolution-based models; our model obtains competitive performance when compared with them. But they are built on neural architecture search [53] that usually requires a considerable amount of compute during the architecture search.

Recent advances in vision-based transformers have achieved great success in image recognition tasks and are comparative with CNN-based models. However, some transformer-based backbones require a considerable number of model parameters with only small improvements in results: ViT-L/16 [51] Swin [136] and DeiT-B [207]. Incorporating convolutions into transformers easily reaches an accuracy of 82%, where the proposed DUCT achieves competitive results over existing work in top-1 accuracy.

Table 5.2 Model performance on downstream tasks (* indicates that the experiments are conducted by ourselves.)

Method	CIFAR 10	CIFAR 100	Pets	Flowers 102
BiT-M [103]	98.91	92.17	94.46	99.30
ViT-B/16 [51]	98.95	91.67	94.43	99.38
ViT-L/16 [51]	99.16	93.44	94.73	99.61
ViT-H/16 [51]	99.27	93.82	94.82	99.51
EfficientNet [202]	98.90	91.7	95.40	98.8
RegNet* [51]	98.7	90.3	93.6	98.9
TNT-B [70]	99.1	91.1	95.0	99.0
CvT [228]	99.16	92.88	94.03	99.62
ours	**99.32**	**94.31**	94.76	99.55

Furthermore, to demystify the trustworthiness of the DUCT decision-making process, we leverage class activation maps to visualize the class responses of the entire DUCT model from output to input [22, 186]. Fig. 5.6 demonstrates that the proposed DUCT can highlight the accurate regions that are highly correlated with ground-truth semantic areas.

Moreover, to investigate the transferability of the pre-trained DUCT model, we also fine-tune and evaluate it on several downstream datasets. Table 5.2 shows that the proposed DUCT can achieve reliable performance on downstream tasks.

Table 5.3 Model performance of semantic segmentation task on ADE20K dataset.

ADE20K		mIoU	#param.
Method	Backbone		
DLab.v3+ [27]	ResNet-101[74]	44.1	63M
ACNet [54]	ResNet-101[74]	45.9	38.5
OCRNet [245]	ResNet-101[74]	45.3	56M
SemanticFPN [102]	ResNet101	38.8	48M
SemanticFPN [102]	PVT [222]	39.8	28M
SemanticFPN [102]	RegNet [222]	35.4	44M
SemanticFPN [102]	EfficientNet [222]	37.1	22M
SemanticFPN [102]	Swin [136]	41.5	32M
SemanticFPN[231]	**DUCT (ours)**	**42.1**	**37M**
UperNet [231]	ResNet-101 [74]	44.9	86M
UperNet [231]	DeiT[207]	44.0	52M
UperNet [231]	Swin[136]	46.1	60M
UperNet[231]	**DUCT (ours)**	**47.2**	**61M**

5.4.3 Image Segmentation

The widely-used semantic segmentation dataset ADE20K [265] is utilized to evaluate the effectiveness of the proposed DUCT backbone. ADE20K contains a total of 25k images that are labeled into 150 semantic categories; 20K of these images are used for training, 2K for validation and the remaining 3K for testing. While there may exist various other semantic segmentation frameworks, our goal is to fairly evaluate the proposed backbone performance. Hence, following the common practice [222, 136], we choose both the semantic-FPN [231, 37] and the UperNet [231] as the segmentation framework, and the model performance is measured by mIoU. AdamW [138] is used for optimization with a linear learning rate scheduler. The initial learning rate is set to 6e-5 with a weight decay of 0.01. The model is trained for 640k iterations with a batch size of 2.

Table 5.3 shows the semantic segmentation results on the ADE20K dataset. In comparing the proposed DUCT backbone with both convolution-based and transformer-based models, we find DUCT has superior performance in terms of mIoU, where it only incurs a slightly higher computational cost than others.

5.4.4 Density Estimation/Regression: Crowd Counting

To further reveal the generalizability of proposed DUCT, we further evaluate it on a density estimation/regression task, namely, crowd counting. Crowd density estimation aims to predict the density map of the number of target objects (e.g., people) in real-world images [196].

The experimental settings are the same as the recent transformer-based model [205] applied to the crowd counting. Benchmark datasets for evaluating the proposed model with DUCT blocks include ShanghaiTech_PartA [258], ShanghaiTech_Part B [258] and UCF_QNRF [88]. The model is trained for 2000 epochs and optimized with the AdamW [138] optimizer. The batch size is set to 4 with learning rate 1e-5. In addition, L2 regularization is adopted as commonly used to avoid overfitting.

Table 5.4 Model performance on crowd counting tasks.

Method	ST_Part A		ST_Part B		UCF_QNRF	
	MAE	MSE	MAE	MSE	MAE	MSE
PACNN [189]	62.4	102.0	7.6	11.8	-	-
S-DCNet [235]	58.3	95.0	6.7	10.7	104.4	176.1
DSSI-Net [129]	60.6	96.0	6.8	10.3	99.1	159.2
BL[143]	62.8	101.8	7.7	12.7	88.7	154.8
RPNet [241]	61.2	96.9	8.1	11.6	-	-
ASNet [93]	57.8	90.1	-	-	91.5	159.7
LibraNet [127]	55.9	97.1	7.3	11.3	88.1	143.7
AMRNet [133]	61.5	98.3	7.0	11.0	86.6	152.2
NoisyCC [214]	61.9	99.6	7.4	11.3	85.5	150.6
DM-Count [214]	59.7	95.7	7.4	11.8	85.6	148.3
GL [215]	61.3	95.4	7.3	11.7	84.3	147.5
SUA-Fully [145]	66.9	125.6	12.3	17.9	119.2	213.3
P2PNet [196]	52.7	85.1	6.3	9.9	85.3	154.5
BCCT [201]	53.1	82.2	7.3	11.3	83.8	143.4
CCTrans [205]	52.3	84.9	6.2	9.9	82.8	142.3
Ours	**52.1**	83.6	**6.1**	**8.6**	**82.1**	**141.5**

As can be seen in Table 5.4, the transformer-based method [205] achieves competitive results compared to the latest work based on convolutional operations. While the proposed DUCT can further improve the estimation of crowd density by properly combining the different feature representations. The qualitative visualizations of the estimated density maps are shown in Fig. 5.8.

5.4.5 Image Retrieval: Person Re-Identification

Image retrieval tasks involve searching for targets (e.g., images) from a gallery to match the query samples. Person re-identification, a prominent task in image retrieval, is considered for evaluating the efficacy of the proposed DUCT block. The experimental setup follows the recent work [75], where the framework built with the DUCT blocks is evaluated on Market1501 [264] and MSM17 [226] datasets.

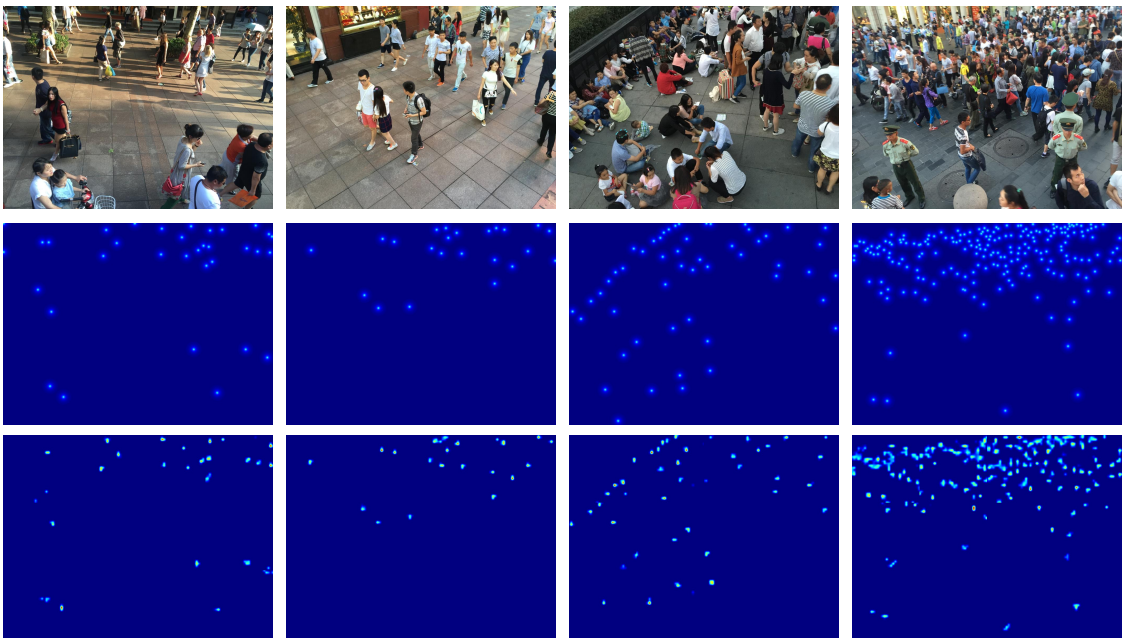


Fig. 5.8 Examples of estimated crowd density maps. From the first row to the last row, they represent the original images, the ground-truth density maps and the estimated density maps as predicted by DUCT.

Table 5.5 shows that recent transformer-based approaches perform slightly better than most CNN-based methods. The proposed backbone also achieves competitively with the latest transformer-based methods. Exemplar retrievals as obtained by the proposed backbone are shown in Fig. 5.9.

5.5 Further Discussion

In this section we discuss some key aspects of DUCT and its impact, primarily based on Fig. 5.10 and Table 5.6, where Fig. 5.10 shows the different levels of information (i.e., local, mid-level and global) learned by our model and Table 5.6 presents the quantitative performance of each of the proposed components.

The impact of Dynamic Local Enhancement Since existing transformers are designed mainly for capturing long-range global information, one of our goals in this work is to enhance the local dependency. In Fig. 5.10 Row-2 blue curves, we observe that our proposed DLE module is able to highlight the potential response that global MHSA otherwise ignored. As the DLE is conducted based on summarizing each token, different local information is re-weighted from each token in a way that is complimentary for global information.

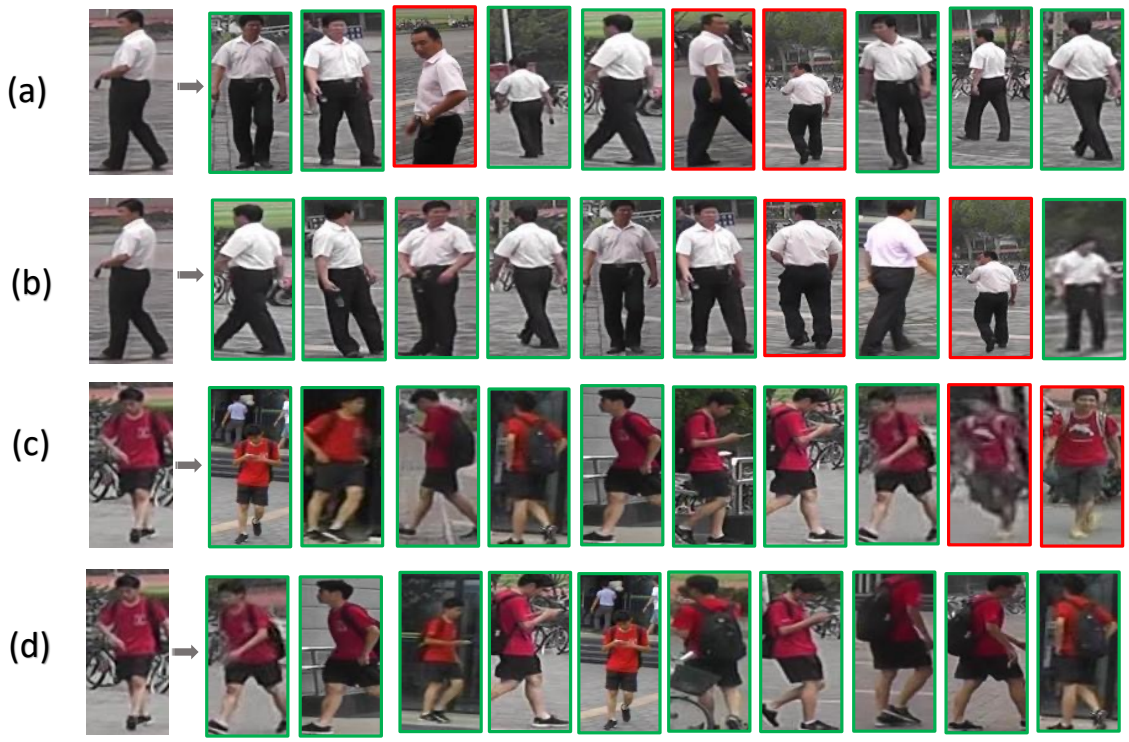


Fig. 5.9 Person retrieval samples from the Market1501 dataset. The first column is the query image, where others are retrieved images from the gallery, which is ranked according to the similarity scores. (a) and (c) are the results based on ViT-B/16. (b) and (d) are the results based on the proposed DUCT. GREEN indicates correctly matched samples and RED indicates mismatched samples.

Table 5.5 Model performance of person re-identification tasks.

Backbone	Method	Market1501		MSMT17	
		mAP	R1	mAP	R1
CNN	CBN [196]	77.3	91.3	42.9	72.8
	OSNet [266]	84.9	94.8	52.9	78.7
	MGN [217]	86.9	95.7	52.1	76.9
	RGA-SC [260]	88.4	96.1	57.5	80.3
	SAN [94]	88.0	96.1	55.7	79.2
	SCSN [30]	88.5	95.7	58.5	83.8
	ABDNet [28]	88.3	95.6	60.8	82.3
	PGFA [146]	76.8	91.2	-	-
	HOReID [216]	84.9	94.2	-	-
	ISP [268]	88.6	95.3	-	-
Transformer	TransReID(DeiT) [75]	88.1	94.9	65.5	83.5
	TransReID(ViT) [75]	88.8	95.0	66.6	84.6
	DUCT(ours)	**89.1**	95.1	**67.4**	**85.9**

Table 5.6 Ablation study of the proposed components on different datasets and different tasks.

MHSA	APM	Dynamic	Unary	cifar100 (acc)	ImgNet (acc)	ADE20K (mIoU)	ST_A (MAE)	ST_B (MAE)	Market1501 (mAP)	MSMT17 (mAP)
✓				90.14	80.3	38.2	57.8	7.4	86.4	61.1
✓	✓			90.73	81.2	39.3	57.0	7.1	87.1	61.9
✓	✓	✓		91.83	81.9	40.7	53.4	6.6	87.6	63.5
✓	✓		✓	93.43	82.4	41.9	54.5	6.9	88.8	65.9
✓	✓	✓	✓	94.31	83.1	42.1	52.1	6.1	89.1	67.4

Also, in Table 5.6, we can see that DLE quantitatively contributes to the final performance improvement.

The impact of Unary Co-occurrence Excitation The UCE module aims to leverage the mid-level information from groups of tokens. Rather than directly model different token information in a dense way, as MHSA, the mid-level information acts as the feature selector (Row-3 in Fig. 5.10) to assign the higher weights for correlated combinations of tokens—which can help discover finer information than dense MHSA from different token groups. The proposed UCE also clearly contributes the final performance improvement as shown in Table 5.6.

Interaction/Limitation of different information levels The class response maps in Fig. 5.10 (Row-1) demonstrate that the proposed backbone can learn to precisely attend to the most relevant regions. Consistently, in the first column, the DLE and MHSA assigned similar

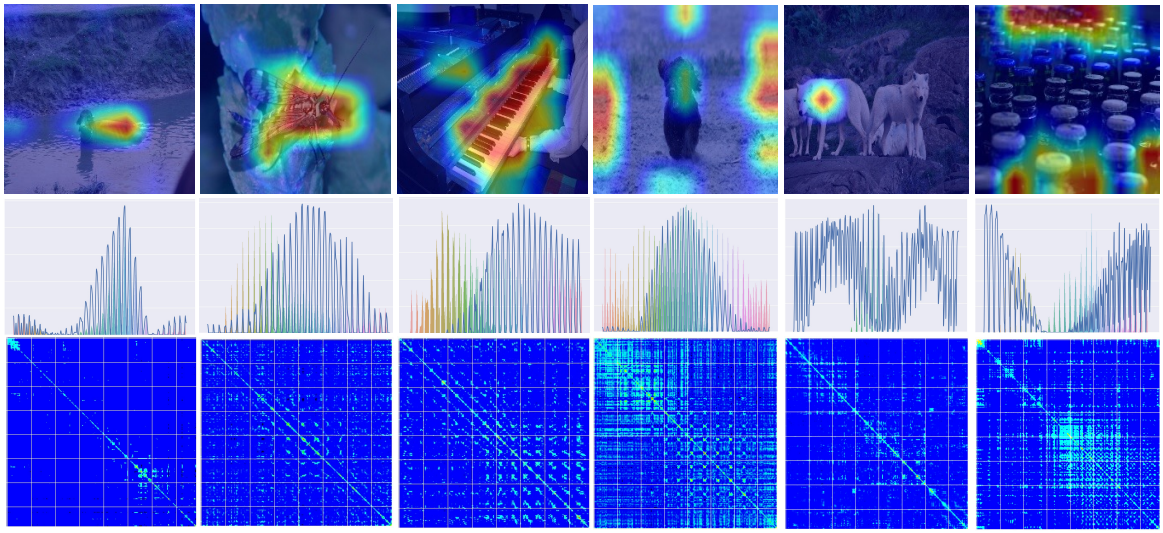


Fig. 5.10 Quantitative analysis of the response values of MHSA global attention, Dynamic Local Enhancement and Unary Co-occurrence Excitation. (Row-1) Visualization of the attention map. (Row-2) Comparison of Dynamic Local Enhancement (DLE) in the blue colour against global attention (MHSA) in the rainbow colour over local tokens. The x-axis is the tokens and the y-axis is the normalized attention response. (Row-3) Visualization of the correlation map in the Unary Co-occurrence Excitation module.

weights on similar tokens, and the UCE unambiguously selects the most informative groups of tokens. A similar observation is also shown in the second column. In the third column, the DLE focuses on smaller regions of tokens since the black-white keys are the most distinct feature to recognize the piano, while the MHSA may further consider other parts of the piano. Also, the UCE still hold the capability to combine the accurate patterns/combination of different token groups. The last three columns show some failure examples, where we can see that there exists an obvious contradiction between DLE and MHSA, which has misguided the model's attention. Such failures are likely caused by a lack of controllable information selection. Since transformer architecture design is still a recent challenge and our main goal in this chapter has been to design a novel parallel transformer architecture, we have directly concatenated the information using simple convolutional layers to select different information as learned by different components, whereas in the future, improvement could likely be made with improved selection of information from different blocks.

Design transformer for diverse vision tasks Existing transformer-based works mainly focus on popular vision tasks (e.g., classification, segmentation). In this chapter, we also benchmarked our model on some other vision tasks like density estimation, image retrieval and downstream task transfer. The results demonstrate the potential of the DUCT framework and the importance of considering different levels of information. In previous years, CNNs

have led progress in various vision tasks, with transformer-based models emerging as a breakthrough due to their expressivity and long-range information handling. Recent deep learning models are graduating towards a unification of various tasks and domains [58]. For the transformer family, this is both an opportunity and a challenge. The information required by different tasks and domains may be highly diverse, where designing a suitable universal model that is able to generalize across these diverse tasks and domains remains an open problem. Our work proposed a parallel hybrid model, paving a novel approach to combine different levels of information into a single model, which could be a reference in future deep network backbone design.

5.6 Conclusion

In conclusion, we have proposed a hybrid transformer named DUCT. This is a parallel structure consisting of Dynamic Local Enhancement (DLE), Unary Co-occurrence Excitation (UCE), and a standard multi-head self-attention module, which together aim to learn the local, mid-level and global information. We found that DUCT outperforms the most recent state-of-the-art approaches on four essential computer vision tasks, i.e., image-based classification, segmentation, retrieval, and density estimation. This work paves a novel way to combine different levels of information, and the results reveal both the viability and validity of the approach. In the future, it would be worth investigating a more controllable selection of different levels of features (e.g., local and global) encoded in a hybrid transformer along with more in-depth theoretical analysis.

Chapter 6

Attention Model for Dynamic Sequential Feature Modelling

Another key data modality rather than the vision signal could be the time series data. Time series data is structured chronological data collected at consistent time intervals and analyzed to uncover trends and patterns over time, whereas vision data consists of unstructured images or videos that represent visual snapshots without a natural sequence. Time series analysis focuses on forecasting, modelling, and prediction using temporal autocorrelation, while computer vision analysis extracts features from images to recognize visual patterns. Time series and vision data require different techniques for processing and analysis based on the structural differences between temporal sequential data versus unstructured pixel data. The earlier chapters highlighted the use of dual attention mechanisms in enhancing computer vision modelling (e.g., image-based crowd scene analysis). In contrast, this chapter successfully applies these mechanisms to time-series data from wearable sensors for human activity recognition. This transition showcases the flexibility and effectiveness of the proposed attention mechanisms across varied data types and domains. It highlights their capability to discern salient features in both spatial and temporal datasets, thereby enhancing model accuracy and efficiency in diverse applications. This shift not only proves the robustness of the attention model but also opens avenues for its application in other areas of deep learning beyond visual and time-series data.

6.1 Introduction

It was widely known that body-worn sensors can be utilised for many real-world applications, including but not limited to sleep monitoring [253], elderly patients assistance [15] and

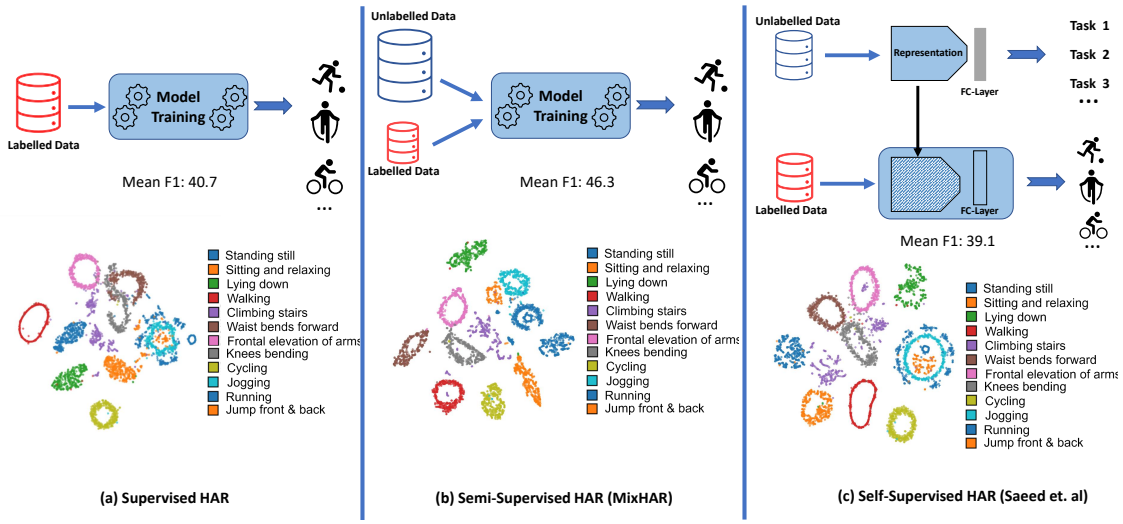


Fig. 6.1 Comparison the feature diagram of different learning paradigm on deep HAR based on same Convolution Neural Network, mHealth dataset and 1% labelled data. The performance of supervised HAR is generally good(a), while the over-fitting still occurred. Self-supervised HAR [181] has its pluses and minuses that may not boost the performance adequately. As we can see, although it improves model's performance on waist bends forward and knees bending, the performance is degraded (with lower mean F1 score) to handle the inter-/intra-activity variability on other activities. Deep semi-supervised HAR (proposed MixHAR) can clearly reduce the intra-activity distance and enlarge the inter-activity distance with better mean F1 score by using unlabelled data. Note in an ideal feature diagram, same class of activities should aggregate to a single point and different class of activities should dispersed as far as possible.

health assessment [152, 56], etc. Out of them, wearable-based human activity recognition (HAR) is one of the core research areas in ubiquitous computing, and it plays an essential role in human behaviour understanding, health monitoring, skill assessment, sports training, etc [26, 65]. In contrast to computer vision-based action recognition, the data was collected by using wearable sensors, making it is less constrained and privacy-friendly[15].

Traditional feature engineering for HAR [169, 15, 86] tends to be a trial-and-error process, which may vary from task to task. Hence, deep learning came into popularity [69] for high-level representations for sensor-based human activities [5, 159, 65]. However, most of the deep learning based HAR rely on the supervised paradigm, which requires substantial labelled data for model training. The unlabelled data in real world scenarios can be collected easily, while the annotations are labour-intensive and time-consuming[15, 108]. Methods that can learn behaviour patterns from unlabelled activities have drawn much attention in the HAR research community, e.g., self-supervised HAR [181, 71], yet effectively learning diverse behaviour patterns from both unlabelled data may not be a trivial task. Fig.6.1

shows the HAR model performance of different learning paradigms, and it seems that the feature learned by a recent self-supervised work [181] [1] on wearable-based dataset does not help much when compared with supervised training. An alternative is deep semi-supervised learning approach, which can learn representation from the labelled and unlabelled data simultaneously [233, 160]. With both labelled and unlabelled data and without the requirement of pre-train tasks design, it may provide a more flexible and reliable solution for HAR, and the HAR model's performance can be boosted by coordinating the limited yet discriminant activities with extra diverse information, in Fig.6.1. (b) we demonstrated an example to show its effectiveness. Although there are some existing semi-supervised HAR works [13, 63, 142, 242], there is a clear gap between the state-of-the-art deep semi-supervised learning and ubiquitous HAR, while one of our goals is to rigorously explore an effective way to take advantage of unlabelled activities in a deep semi-supervised fashion for ubiquitous HAR (Deep-Semi-HAR).

In this chapter, we summarized and defined a deep semi-supervised HAR research pipeline, then we developed and evaluated five conventional/popular deep semi-supervised techniques on existing HAR works [158]. Most of them focus on regularising the consistency in different perturbed unlabelled data while may suffer from the non-HAR-specialise perturbations. Also, such consistency regularisation help to handle the intra-activities variability while may ignore the inter-activities variability. Motivated by these observations and to better take advantage of wearable-based unlabelled activities, we proposed a deep semi-supervised HAR approach named MixHAR. MixHAR, which is motivated by the recent stateof-the-art semi-supervised approaches in other research domains (e.g.,MixMatch [12]), aims to leverage the unlabelled data to train a robust model by mixing labelled activities and unlabelled activities simultaneously with a linear interpolation, which also considered the inter-/intra-activities variability. The transitions between the labelled/unlabelled data incentivising the robust network training smoothly. Although the linear interpolation technology is widely used in different research tasks[195, 24, 246], directly utilizing them with HAR may not be appropriate since the sensor-based human activities hold the unique characteristics. We firstly seek an appropriate way to mix the labelled and unlabelled activities in deep semi-supervised fashion, then we located a problem during the activities mixing that there are some conflicts between the mixed activities and original activities (activity-intrusion problem), which is crucial for the completeness of our MixHAR. Therefore, a mixing calibration mechanism is proposed to alleviate the activity-intrusion problem on the feature embedding space. The final MixHAR achieved significant performance and shows the effectiveness of using deep

[1] Both [181] and [203] are the methods that aim to take advantage of unlabelled data. For clearly and fairly comparison, we choose the [181]

semi-supervised techniques to take advantage of unlabelled activities. Our main contribution of this chapter can be summarised as follow.

Contribution

- Borrowing the recent advancements of semi-supervised techniques from deep learning, we re-produced and evaluated five conventional/popular deep semi-supervised techniques on wearable-based HAR. Although they may not always improve the deep HAR model's generalisability, results still showed the feasibility of these five Deep-Semi-HAR methods to make use of unlabelled activities, and they could be further improved, which paves the way in future deep semi-supervised HAR research.
- On top of these conventional/popular Deep-Semi-HAR, we proposed a stronger Deep-Semi-HAR framework named MixHAR. MixHAR firstly generated the pseudo labels for unlabelled activities with entropy minimisation. Then labelled/unlabelled activities were mixed by linear interpolation to train the model simultaneously. The proposed MixHAR could better take advantage of unlabelled activities with handling diverse training variability (inter/intra-activities variability), which improved the discrimination and generalisation of deep HAR model.
- The mixing/interpolation suffered from the conflicts between the mixed activities and original activities, which we identified as the activity-intrusion problem. Here we proposed a calibration attention mechanism by exploring the correlation on feature space for mixed activities to alleviate the activity-intrusion problem, which further enhanced the model's capability of handling the substantial variability from mixed activities.
- The results showed our MixHAR outperformed both the five conventional/popular Deep-Semi-HAR and other recent methods that also take advantage of unlabelled data (e.g., both self-supervised HAR and semi-supervised HAR). We believe this chapter could be a good reference for future semi-supervised HAR research.

6.2 Related Work

This work aims to alleviate the training over-fitting of deep HAR, and we take advantage of unlabelled data along with limited labelled data simultaneously in a deep semi-supervised fashion. This section will cover the algorithmic background to establish the context for both deep HAR and deep semi-supervised learning.

6.2.1 Human Activity Recognition

Human activity recognition (HAR) can be divided but not limited to two main research directions: vision-based HAR and sensor-based HAR [5]. For vision-based HAR, the data are substantially recorded as video frames, skeleton modality, etc [267, 77]. Earlier approaches proposed to map the information of activities into 1D hand-crafted features, representing some points of interest that may significantly change in temporal and spatial space [45]. Recent research focused on designing the deep learning models to automatically extract the spatial-temporal features, such as two-stream CNN [191], LSTM-based models and 3D-CNN-based models[91, 198].

For sensor-based HAR, compared with vision-based HAR, it can preserve the individual's privacy, efficiently saving the computational cost, etc [5]. Using wearable/mobile sensing data for human activity recognition has a long-standing history in the ubiquitous computing community. Early sensor-based HAR approaches tried to utilise the single accelerometer signals to recognise the human activities [9, 107]. With the great success of feature extraction and variability handling, deep learning has been the mainstream to extract the appropriate features in the human activity recognition system. Guan and Plötz proposed a simple yet powerful deep ensemble framework to combine strong LSTMs with low bias and high variance into a robust learner with variance reduction [65]. Most recently, a discriminative adversarial multi-view network [5] was proposed to model the multi-modal spatial-temporal patterns from the sensory data. This chapter proposed a simple but effective Deep-Semi-HAR framework to alleviate the training over-fitting on limited labelled activities by taking advantage of extra unlabelled data.

6.2.2 Deep Semi-Supervised Learning

Deep learning methods provide remarkable improvement for pattern recognition, yet the lack of training data may cause the over-fitting problem. Although extensively collected more data may improve the models' generalisation ability, the annotation can be expensive. One the other hand, unlabelled data can be easily acquired, which can be employed for representation learning. Deep semi-supervised learning is a deep learning method between supervised learning and unsupervised learning, which is trained simultaneously with the data that has a few labels but is mostly unlabelled. Deep semi-supervised learning can be used if the data has any of the following three main assumptions [18]. Firstly, if data points lie in the same cluster in feature space, they should likely be the same class. Secondly, if data points belong to the same classes or clusters, their outputs from the deep model should be close. The third one is that the high-dimensional features of data should roughly

lie on a low-dimensional manifold, and the classification boundary should not across the high-density regions. These three assumptions are suitable for most standard classification tasks, which should be also suitable for the sensor-based human activity recognition [242, 65]. Based on these assumptions, it is expected that if a perturbation is applied to the unlabelled data points, the corresponding outputs should still lie very closely, which is formed as consistency regularisation that most semi-supervised methods rely on [158]. Furthermore, proxy-label [114] method generated pseudo labels for unlabelled data as additional training examples based on some heuristic. Generative models [20] also can be used to learn the data distributions from massive unlabelled data and transfer to downstream tasks with limited labelled data [71], which is also known as self-supervised learning or transfer learning.

6.3 Methodology

Fig.6.2 shows the overview of our methods, which aims to take advantage of substantial unlabelled activities in semi-supervised fashion.

6.3.1 Deep-Semi-HAR Pipeline

In supervised HAR [69, 65, 109], the training data consists of the time series sensor readings of different activities as $\{\mathbf{X}, \mathbf{y}\}$, where $\mathbf{X} \in \mathbb{R}^{H \times T}$ and $\mathbf{y} \in \{1, 2, ..., C\}^T$, H is the number of sensor channels, T denotes the number of temporal samples and C is the total number of classes. Following the previous work in HAR [15], the data are segmented into N frames along each sensor channel by a sliding window (with fixed length) and overlapping rate. Then we can obtain the frame-wise time series data $\mathscr{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^N$, where $\mathbf{X}_i \in \mathbb{R}^{H \times L}$, and the label $y_i \in \{1, 2, ..., C\}$, where L is the length of sliding window. The testing data is processed with same operation. The deep HAR model with activity representation learning module $\mathscr{F}$ and classification module $\mathscr{S}$ can be end-to-end trained by a loss function (e.g., Cross Entropy Loss) using training data $\mathscr{D}$, and evaluated on unseen testing data.

Then we define the protocol to split the labelled/unlabelled data for Deep-Semi-HAR (including our proposed MixHAR). The training data $\{\mathbf{X}, \mathbf{y}\}$ can be split into two parts, namely labelled part $\{\mathbf{X}^\ell, \mathbf{y}^\ell\}$ and unlabelled part $\{\mathbf{X}^u, \mathbf{y}^u\}$. A certain percentage (e.g., 1%,3%, etc.) of entire training data are used as labelled data and the rest is utilised as unlabelled data. More specifically, the percentage partition is applied to each class. For instance, each class in the labelled data pool consists of a certain number of activities from the corresponding class in the entire training data. Therefore, the class balance/imbalance distributions in both labelled and unlabelled parts are the same as the original training

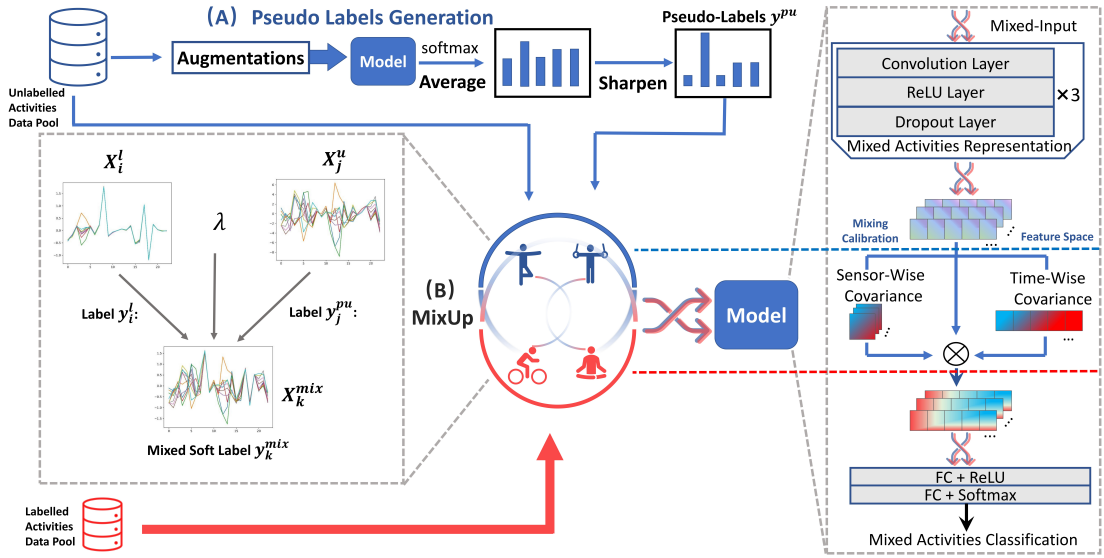


Fig. 6.2 Overview of our proposed method. We first generate the pseudo labels for unlabelled activities and then mix them with labelled activities. The mixed activities are utilised to train the model directly. A mixing calibration mechanism is applied inside the feature space of mixed samples and placed between the representation learning module and classification module. The $\mathbf{X}_i^\ell$ denote a example of labelled activities with corresponding label $\mathbf{y}_i^\ell$. The $\mathbf{X}_j^u$ denote a example of unlabelled activities and $\mathbf{y}_j^{pu}$ is the corresponding pseudo label. The $\mathbf{y}_k^{mix}$ and $\mathbf{X}_k^{mix}$ denote a example of mixed activities.The sensor data from different activities are mixed using linear interpolation. This involves combining pairs of signals and their corresponding labels to a varying degree, determined by a coefficient typically drawn from a Beta distribution.

dataset. To avoid the information leaking, labelled/unlabelled data partition is applied BEFORE the aforementioned data segmentation, which leads to the labelled activity data $\mathscr{D}^\ell = \left\{(\mathbf{X}_j^\ell, y_j^\ell)\right\}_{j=1}^{N^\ell}$ and unlabelled activity data $\mathscr{D}^u = \left\{(\mathbf{X}_k^u)\right\}_{k=1}^{N^u}$. We set $N^u >> N^\ell$ [158, 220] to mimic the practical semi-supervised HAR scenario. In supervised baseline, deep HAR model training is only relied on the labelled data $\mathscr{D}^\ell$. Keeping utilising the labelled data $\mathscr{D}^\ell$, Deep-Semi-HAR also makes use of the unlabelled data $\mathscr{D}^u$, specifically the labelled and unlabelled data are used to train the model SIMULTANEOUSLY in an end-to-end fashion, which aims to alleviate the over-fitting and improve the model's generalisation (one example is shown in Fig.6.1).

6.3.2 Labelled and Unlabelled Activities Mixing

Following the deep semi-supervised fashion, the labelled and unlabelled data are leveraged to train the model simultaneously in our framework. Our approach is motivated by previous

work MixUp [255] that is to mix two samples by linear interpolation, and we extend the MixUp to interpolate the information between existing limited labelled and extra unlabelled activities. MixUp can be viewed as a novel data augmentation in this work and we leverage the mixed new data (activities), which may not only contain discriminant classification information but also inject diverse variability, to regularise the model learning to be more robust. In deep HAR, a more robust model can be obtained by handling the inter/intra-activity variability [242]. Inter-activity variability occurs when the activities belong to different class but show similar characteristics (e.g., knees bending and climbing stairs). Intra-activity variability occurs when the same activities being performed by different subjects, or even by the same subject but within the influence of emotional or environmental factors [15].

When mixing the unlabelled activities with labelled activities together in MixHAR, there are two cases that may lead deep HAR model to be more robust : 1) if labelled and unlabelled activities are mixed and they belong to different activities, the model's generalisation can be improved by training with injecting diverse inter-activity variability. 2) if labelled and unlabelled activities are mixed and they belong to same activities but show different patterns, the model's generalisation can be improved by training with injecting diverse intra-activity variability. Furthermore, by means of the interpolation between labelled and unlabelled data, the network can be incentivised to learn the transitions between each pair-wise samples (that used to mix). The transitions connect the existing supervised classification information from labelled data and more diverse activities variability from unlabelled data. Rather than using labelled data and unlabelled data in a separate way, our approach helps deep HAR model to learn the discriminant patterns of different activities with handling large variability simultaneously.

Given a batch (batch size is B) of labelled data $\mathscr{B} = \left\{(\mathbf{X}_j^\ell, y_j^\ell)\right\}_{j=1}^{B}$, $\mathscr{B} \in \mathscr{D}^\ell$ and a same sized batch of unlabelled data $\mathscr{U} = \left\{(\mathbf{X}_k^u)\right\}_{k=1}^{B}$, $\mathscr{U} \in \mathscr{D}^u$, we firstly generate the pseudo labels for unlabelled data (Fig.6.2.(A)). The pseudo labels should be at least more accurate than a manually defined classification thresholds [194] and give a relative clear guidance to model when it trained with different variability. Motivated by the previous work Unsupervised Data Augmentation [232] that leveraged the multiple augmentations to inject noisy variability and obtained robust pseudo labels, two data augmentation functions $\mathscr{A}_1$ and $\mathscr{A}_2$ are applied on unlabelled data [194]. For each[1] unlabelled data $\mathbf{X}_k^u$, the initial pseudo label is firstly generated with deep model $\mathscr{F}$, which is formed as:

$$\mathbf{y}_k^{pu} = \frac{1}{2}(\text{softmax}(\mathscr{F}(\mathscr{A}^1(\mathbf{X}_k^u))) + \text{softmax}(\mathscr{F}(\mathscr{A}^2(\mathbf{X}_k^u)))) \tag{6.1}$$

[1]Note for demonstration, we only use one sample as example

here the $\mathbf{y}_k^{pu}$ is a categorical probability vector. The HAR model is expected to predict consistent labels for the same activities with different variability, hence we use average on all predictions rather than any single predictions from activity sample. There exists a large number of diversities/variabilities in unlabelled activities (50%-99% of training data are unlabelled). Without the real guidance/labels, the model's predictions (i.e., pseudo labels) for the unlabelled activities may contain large uncertainties. These uncertainties may lead to ambiguous classification boundaries [209]. Consequently, the class probabilities (the prediction normalised by softmax function) may close to a uniform style (as shown in Fig.6.2.(A) and visualisation in Section.5.3) with relatively high entropy. Such uniform distribution may not provide the discriminative classification information, while we want the pseudo labels can at least represent different activities with proper discrimination. Motivated by the entropy minimization in deep semi-supervised learning [150, 62], one way to alleviate the uncertainties in prediction is to enforce model output low entropy predictions on unlabelled data. In this work, we apply an sharpen operation to the categorical distribution to reduce the uncertainty for each $\mathbf{y}_k^{pu}$, which is formed as:

$$\text{Sharpen}\left(\mathbf{y}_k^{pu}, v\right) = \frac{\left(\mathbf{y}_k^{pu}\right)^{\frac{1}{v}}}{\left\| \left(\mathbf{y}_k^{pu}\right)^{\frac{1}{v}} \right\|_1} \tag{6.2}$$

here the output will approach 'one-hot' distribution when $v \to 0$. After sharpen operation, the pseudo labels may contain more discriminative activity classification information. The unlabelled data batch with pseudo labels can be obtained as $\mathscr{U} = \left\{(\mathbf{X}_k^u, \mathbf{y}_k^{pu})\right\}_{k=1}^{B}$. Also, the labels of annotated data are changed to be the 'one-hot' encoding such that $\mathscr{B} = \left\{(\mathbf{X}_j^\ell, \mathbf{y}_j^\ell)\right\}_{j=1}^{B}$. Then unlabelled data with two augmentation functions give us $\mathscr{U}^{A^1} = \left\{(\mathscr{A}^1(\mathbf{X}_k^u), \mathbf{y}_k^{pu})\right\}_{k=1}^{B}$ and $\mathscr{U}^{A^2} = \left\{(\mathscr{A}^2(\mathbf{X}_k^u), \mathbf{y}_k^{pu})\right\}_{k=1}^{B}$. A super batch can be built as $\mathscr{X} = \mathscr{B} \cup \mathscr{U}^{A^1} \cup \mathscr{U}^{A^2}$, where $\mathscr{X} = \left\{(\mathbf{X}_k^s, \mathbf{y}_k^s)\right\}_{k=1}^{3B}$. During the training, given the interpolation (i.e., MixUp) parameter $\lambda \sim \text{Beta}(\beta, \beta)$ and $\lambda = \max(\lambda, 1-\lambda)$, each paired data points $(\mathbf{X}_k^s, \mathbf{y}_k^s)$ from $\mathscr{X}$ and $(\mathbf{X}'^s_k, \mathbf{y}'^s_k)$ from *Shuffled* $\mathscr{X}$ are mixed (Fig.6.2.(B)) to be a single data points is following:

$$\mathbf{X}_k^{mix} = \lambda \mathbf{X}_k^s + (1-\lambda)\mathbf{X}'^s_k, \quad \mathbf{y}_k^{mix} = \lambda \mathbf{y}_k^s + (1-\lambda)\mathbf{y}'^s_k \tag{6.3}$$

where we can obtain the virtual mixed data batch $\mathscr{X}^{mix} = \left\{(\mathbf{X}_k^{mix}, \mathbf{y}_k^{mix})\right\}_{k=1}^{3B}$. The virtual mixed activity data are directly used to train the deep HAR model.

6.3.3 Activity Intrusion and Mixing Calibration

Although leveraging the mixed activities helps the HAR model to obtain better generalisation, not all the mixed activities have the positive impact, when the conflicts occur between the mixed data and original data, yielding the biased HAR model training. We defined this problem as activity-intrusion problem [66]. Intuitively, the activity-intrusion problem occurs when any virtual mixed activity is similar to a specific real activity but is assigned with different virtual mixed soft label distribution. For instance, if a labelled activity is walking and an unlabelled activity is running, the resulting mixed activity may very likely be similar to jogging, but the soft label of the mixed activity is mainly assigned as the probabilities of "walking" and "running". Such activity-intrusion problem may result in degradation of the model's discrimination/performance, and affect the learning granularity. Here motivated by previous work [66] that further feature correlation exploration may alleviate the intrusion problem when a model is trained with MixUped samples, thus we propose a mixing calibration mechanism that learns the feature correlation inside the embedding space, where the correlation is the high order statistics (e.g., covariance) [89, 118, 206, 52] derived from the output features, which also reveals the relationship between the pair-wise mixed activities. Then the obtained correlation is formed as an attention guiding model for more discriminant representations. Each virtual mixed activity sample is input to the representation learning module to obtain the activity representation $\mathbf{F}_k$, where $\mathbf{F}_k = \mathscr{F}(\mathbf{X}_k^{mix})$ [1] and $\mathbf{F}_k \in \mathbb{R}^{c\times t}$, the c dimensions contains the multi-sensor information and the t dimensions contains the temporal information, and we compute both sensor-wise and time-wise feature covariance matrix, respectively $\mathbf{M}_c$ and $\mathbf{M}_t$, where they can be calculated as:

$$\mathbf{M}_c = \mathbf{F}_k\, \bar{\mathbf{I}}_c\, \mathbf{F}_k^T, \qquad \mathbf{M}_t = \mathbf{F}_k^T\, \bar{\mathbf{I}}_t\, \mathbf{F}_k, \tag{6.4}$$

where $\bar{I}_c = \frac{1}{t\times 1}(\mathbf{I}_c - \frac{1}{t\times 1}\mathbf{1}_c)$, $\mathbf{I}_c \in \mathbb{R}^{t\times t}$, $\mathbf{1}_c \in \mathbb{R}^{t\times t}$ and $\bar{\mathbf{I}}_t = \frac{1}{t\times 1}(\mathbf{I}_t - \frac{1}{t\times 1}\mathbf{1}_t)$, $\mathbf{I}_t \in \mathbb{R}^{c\times c}$, $\mathbf{1}_t \in \mathbb{R}^{c\times c}$. The obtained covariance matrix $\mathbf{M}_c \in \mathbb{R}^{c\times c}$ and $\mathbf{M}_t \in \mathbb{R}^{t\times t}$ have clear structural information that each row or column [2] contains statistical dependency of the features along the sensors and times, which also indicates the feature correlation of each pair of activities in mixed sample. To learn such structural correlation information in end-to-end manner, we apply a group convolution [106] directly on obtained covariance matrix by utilising independent convolution filters for each row or column (here we apply it on each row) [52, 36, 80] . The filter size, number of filters and number of groups [3] are all set as the c for $\mathbf{M}_c$ or t for $\mathbf{M}_t$

[1] Note for demonstration, we only use one sample as example

[2] The covariance matrix is diagonally symmetrical

[3] parameters of convolution in Pytorch

[52, 57],so $\mathbf{M}_c \in \mathbb{R}^{c\times c}$ and $\mathbf{M}_t \in \mathbb{R}^{t\times t}$ can be derived into $\mathbf{M}_c \in \mathbb{R}^{c\times 1}$ and $\mathbf{M}_t \in \mathbb{R}^{t\times 1}$. Then one standard convolution (filter size is 1 and number of filters are c or t) is applied to enhance the correlation learning, and the resulting $\mathbf{M}_c \in \mathbb{R}^{c\times 1}$ and $\mathbf{M}_t \in \mathbb{R}^{t\times 1}$ are activated by sigmoid function, which acts as the attention weight vector. Then we reshape the $\mathbf{M}_t \in \mathbb{R}^{t\times 1}$ to be $\mathbf{M}_t \in \mathbb{R}^{1\times t}$, and the mixing calibration attention can be formed as $\mathbf{M}_c \times \mathbf{M}_t$, and final activity representation can be calculated as $\mathbf{F}_k \otimes (\mathbf{M}_c \times \mathbf{M}_t)$, where the $\otimes$ denotes the multiplication operation. The final representation is input to classification module $\mathscr{S}$ to make the prediction. Mixing calibration can alleviate the activity-intrusion problem and further enhance/rectify the model's discrimination and improve the model's capability of handling the substantial injected/mixed variability from virtual mixed activities.

Finally in each training iteration, if $\mathbf{X}_k^s \in \mathscr{B}$ or $\mathbf{X}'^s_k \in \mathscr{B}$, we calculate the cross entropy loss as a lot of information comes from labelled data with smoothly injecting the information from unlabelled data. If $\mathbf{X}_k^s \notin \mathscr{B}$ and $\mathbf{X}'^s_k \notin \mathscr{B}$, we calculate the mean square error (with a loss weight γ) as HAR models are mainly learned from unlabelled data with pseudo labels and mean square error is less sensitive to outliers in predictions.

6.4 Experiments

6.4.1 Datasets

Following previous works [5, 69, 44], our experimental evaluation is conducted on five (4+1) benchmark datasets, which correspond to diverse yet typical applications in the field of wearable-based HAR, namely the *Oppotunity* [19], the *PAMAP2* [177], the *mHealth* [8], the *DSADS* [10]. Also, on top of the mHealth dataset, the mHealth+ dataset further contains the imbalanced NULL-Class Activity (more than 70% of entire dataset). Other settings in the mHealth+ dataset are same as mHealth dataset. Our goal is to fairly show an evaluation of NULL-Class [26] impact between mHealth and mHealth+ datasets.

Following the previous works, we set a sliding window of 1 second with 50% overlapping for Opportunity dataset [69], a small window size of 0.8 seconds and 50% overlapping for DSADS datset [5], and a large window size of 168 time points and 78% overlapping is applied to others[69, 188].

6.4.2 Evaluation Protocol

Following the previous works [242, 25, 69], we apply the Leave One Subject Out Cross Validation (LOSO-CV) as evaluation strategy, where the test data is the activities from the leave one out subject, and the final results is the average (with the standard deviation) over

iterating all the subjects. Also, we use the mean F1 score (F_m) as evaluation metrics to measure the performance of different methods, which is calculated as:

$$F_m = \frac{1}{C}\sum_{c=1}^{C}\frac{2\mathrm{TP}_c}{2\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}, \tag{6.5}$$

where C is the total number of class, TP_c is the true positive of each class, FP_c is the false positive of each class, FN_c is the false negative of each class.

6.4.3 Deep-Semi-HAR Baselines

Recently, deep semi-supervised learning has achieved state-of-the-art in different research areas [160, 194, 209]. Most of them are based on five conventional/popular works [160, 158]. To bridge the gap between recent deep semi-supervised learning and ubiquitous HAR with considering the compatibility, we reproduced and developed the five conventional methods to existing HAR works. These methods solely involved predicting pseudo labels for unlabelled data or conducting regularization terms using unlabelled data. The comprehensive description/analysis of each method is out of the scope of this chapter and we refer interested readers to the original works [111, 150, 114, 204, 62]. Instead, in this chapter, we aim to apply and evaluate these five conventional deep semi-supervised techniques on HAR and we followed the original works to conduct the experiments. Most importantly, some of these methods may be sensitive to the training hyper-parameters, where we carefully tuned them and list the values/settings here for reproducibility of our results.

π-Model [111] is a training framework with two terms of losses. One loss is the standard Cross-Entropy loss based on the limited labelled activities, another one conducts the consistency regularization between model predictions on unlabelled activities and the model predictions on augmented unlabelled activities. The consistency regularization in this approach may help the model to be more robust to variability in same activity. The data augmentation used here is time series scaling, the consistency regularization loss is based on the mean square error with the weight γ. Different augmentations may affect the model performance, while the HAR-specific augmentation may be explored in future work. Instead of conducting the consistency regularization with randomly chosen augmentation in π-Model, Virtual Adversarial Training[150] (**VAT**) automatically approximates a perturbation for each unlabelled activity in an adversarial direction. The adversarial direction is the direction that the label probabilities are most sensitive in input space. Intuitively, this adversarial approximation helps the model to learn the invariant characteristic from the activities with different variability. In original works, there are three hyper-parameters that may affect the model performance. Motivated by original works [150], in our experiments for VAT, the

number of iterations to find the adversarial direction is set to 1, the perturbation size for adversarial direction is set as 10 and the regularization coefficient that controls the output adversarial value is set as 0.8, and a weight for consistency regularization term on unlabeled data is set as γ. Entropy Minimization [62] is the method that add a loss term into standard deep semi-supervised method that encourages the low-entropy prediction from deep model. We follow the previous work to use the optimal solution that combine it with VAT as **VATENT** [150]. The entropy minimization loss is directly added to the VAT loss group without weight control. Rather than directly using the same model on both labelled and unlabelled data, in Mean-Teacher[204] (**MT**), an extra model is derived from the existing model for labelled data by applying the Exponential Moving Average technique, and the consistency regularization is conducted on the outputs between two models with same unlabelled activities. We set the exponential moving average decay as 0.9 and weight of consistency regularization as γ. Furthermore, Pseudo-Labeling [114] (**PL**) aims to produce the pseudo labels for unlabelled activities using the model itself over the iteration of training. Hence model's generalisation may be improved with more diverse training data. We also compare our method with the state-of-the-art **MixMatch** work [12], which is most similar to our proposed works. The proposed MixHAR (w/o Cal) is the version that is carefully adapted from the MixMatch for human activity recognition, hence we did not additionally list the mixmatch in our comparison. In each epoch, the model is firstly trained with labelled activities, and then it predicts the pseudo labels for unlabelled activities. The labelled activities with labels and unlabelled activities with pseudo labels are used all together to optimize the model (with cross-entropy loss) in the next epoch under the standard supervised training scheme. The model architecture used here is the same as the model used both in the supervised baseline and our MixHAR. Following the most deep semi-supervised works [160, 158], the weight γ follows the ramp-up strategy from zero. They are all trained 150 epochs with learning rate 10^{-3} and Adam optimizer.

6.4.4 Implementation Details

We conducted our experiments by using Python and Pytorch on Ubuntu platform with NVIDIA RTX TITAN GPU. Motivated by recent works [181, 239, 69], there are 3 activity representation encoding blocks and two fully connected layers used as activity classification modules in our HAR model. Each block that encoding the representation consists of a single 1D convolution layer with ReLU activation and dropout techniques (with rate 0.3). The number of convolution filters is set as 32, 64, 96 and filter sizes are 5, 5 and 3 for convolution in these 3 blocks [239]. The output from the first fully connected layers consists of 64 units and is activated by the ReLU function, the output from the second fully connected layer

consists of the units that reflect the number of classes, which is input to a softmax function to obtain the classification probability. The mixing calibration is placed before the activity classification module and the details of dynamic calculation were presented in Section.3.3. We followed the standard mini-batch training with setting batch size 32, small batch size is used as the amount of labelled data are small. In each iteration, we sampled two batches of activity frames (with the same batch size), one batch from the labelled data pool and another batch from the unlabelled data pool. Since the total number of labelled data are far less than the total number of unlabelled data, we followed the previous work [220, 150, 149, 232], that we cyclically sampled the labelled batches until all the unlabelled data are used once. With Leave-One-Subject-Out Cross Validation (LOSO-CV), the deep network is trained 150 epochs by Adam [99] gradient decent optimizer with learning rate 10^{-3}. We use the Re-scaling [208] and TimeWarping[208] augmentation for the pseudo-label generation parts, and the β is set as 0.8 for data mixing, v is set as 0.4 for sharpen operation. Following previous works [204, 111, 150, 114, 160], we set the aforementioned loss weight γ in MixHAR and Deep-Semi-HAR baselines by following a ramp-up training strategy [158, 160].

6.5 Results and Discussion

In this section, we provide in-depth discussion based on the performance of our MixHAR.

6.5.1 Comparison of Different Deep-Semi-HAR

Table.6.1 shows the comparison of different Deep-semi-HAR performance. We summarised the vital observation as follows.

- The model in supervised baseline only trained with labelled data. Comparing with the supervised baseline, the five conventional deep semi-supervised techniques helped the model training in some labelled/unlabelled data settings, specifically on the dataset with repetitive activities and without NULL-Class activities such as DSADS, mHealth. Nevertheless, the performance of these five methods was not always satisfactory, and the possible reason for such observation may be related to the rationale behind, where these methods mostly focused on alleviating the intra-activity variability. When with less labelled data, the primary target may be the understanding of what the current activity is, the inter-activity variability may be more crucial for model training. For example, in the DSADS dataset, the users are asked to freely perform different activities in their own style [10], which results in relatively larger intra-activities variability, so these methods obtained obvious effect on the DSADS dataset. In the Opportunity

Table 6.1 Comparison of different HAR approaches on different dataset, the percentage denotes the amount of labelled data partitioned from the training data. The bold highlight the performance (F_m) of MixHAR, which obtained best performance.

		0.5%	1%	3%	5%	10%	30%	50%
PAMAP2	Supervised	26.7(8.2)	34.0(8.5)	39.9(12.7)	43.8(11.9)	44.1(13.6)	74.2(17.7)	82.5(14.5)
	MT-HAR	17.5(9.3)	28.0(8.9)	33.5(13.3)	42.5(13.6)	46.2(16.3)	74.6(16.5)	82.9(14.6)
	Pi-HAR	27.4(11.1)	36.0(8.4)	46.4(11.4)	48.5(8.9)	50.4(15.7)	74.4(14.2)	80.6(16.7)
	PL-HAR	27.1(7,9)	27.0(8.5)	31.2(10.3)	34.8(12.9)	38.9(11.8)	76.4(14.0)	78.9(16.4)
	VAT-HAR	20.5(7.1)	25.3(7.9)	34.7(12.6)	35.6(13.6)	47.9(16.2)	80.3(16.3)	82.2(15.4)
	VATENT-HAR	18.5(7.8)	21.4(6.6)	20.2(14.3)	24.9(12.9)	34.7(15.2)	66.1(15.4)	70.3(20.1)
	MixHAR(w/o Cal)	28.0(9.3)	37.0(8.5)	42.3(10.2)	49.9(11.1)	53.4(13.3)	76.0(12.1)	82.7(15.4)
	MixHAR	**28.1(6.7)**	**38.7(6.7)**	**43.4(12.1)**	**48.9(11.2)**	**55.2(14.2)**	**80.7(12.2)**	**83.3(12.8)**
mHealth	Supervised	38.5(8.0)	40.7(10.6)	43.0(13.6)	44.6(8.2)	42.3(12.3)	70.4(15.5)	86.0(9.3)
	MT-HAR	37.8(6.5)	43.1(7.5)	44.0(13.0)	45.6(8.9)	39.5(10.1)	69.7(16.8)	81.3(7.0)
	Pi-HAR	36.7(6.9)	40.3(8.1)	45.6(11.6)	43.1(7.9)	37.5(13.0)	71.5(14.3)	82.8(9.7)
	PL-HAR	39.2(7.0)	41.8(7.1)	46.3(11.2)	46.7(6.8)	42.0(10.2)	73.1(13.6)	83.7(7.4)
	VAT-HAR	38.8(6.5)	40.3(8.0)	45.9(10.1)	44.0(7.0)	39.1(11.4)	72.8(13.5)	80.1(9.9)
	VATENT-HAR	40.0(7.3)	39.4(7.4)	43.0(10.5)	46.6(7.6)	46.6(6.9)	76.0(11.3)	83.2(9.9)
	MixHAR(w/o Cal)	40.1(7.8)	44.8(10.3)	51.4(12.1)	53.2(7.9)	51.5(9.8)	82.3(11.5)	91.4(5.1)
	MixHAR	**43.8(7.6)**	**46.3(10.3)**	**52.5(11.8)**	**53.8(7.5)**	**55.1(10.1)**	**82.3(10.4)**	**92.3(4.3)**
DSADS	Supervised	41.0(4.9)	50.6(20.4)	59.5(8.6)	61.0(7.4)	62.8(5.8)	84.5(7.0)	86.7(6.3)
	MT-HAR	48.2(2.3)	56.5(4.8)	59.4(6.8)	62.9(7.1)	61.8(5.6)	83.9(6.7)	86.0(7.3)
	Pi-HAR	49.2(4.1)	57.3(6.5)	61.4(7.8)	60.2(7.3)	63.0(8.0)	82.2(8.8)	85.8(9.2)
	PL-HAR	47.1(3.0)	58.0(5.8)	63.6(9.3)	64.4(4.4)	65.3(6.9)	81.6(10.1)	86.1(7.3)
	VAT-HAR	49.1(2.2)	56.3(7.0)	58.9(7.3)	64.2(3.4)	64.4(8.4)	81.6(8.1)	84.9(7.6)
	VATENT-HAR	46.8(3.1)	56.2(5.7)	58.1(7.6)	60.0(7.6)	61.0(7.2)	87.3(6.6)	87.1(8.0)
	MixHAR(w/o Cal)	44.5(2.1)	58.7(6.6)	64.6(6.6)	63.1(6.9)	66.1(7.2)	85.8(5.2)	89.1(4.3)
	MixHAR	**50.3(4.6)**	**59.5(6.3)**	**66.6(8.4)**	**64.9(5.3)**	**67.8(6.5)**	**89.9(6.9)**	**90.1(7.2)**
mHealth+	Supervised	13.2(4.1)	19.6(6.2)	25.0(8.5)	26.5(9.6)	25.8(9.5)	44.7(14.9)	52.5(11.0)
	MT-HAR	14.2(3.4)	24.1(5.0)	25.2(7.0)	23.9(7.1)	20.1(8.5)	44.8(13.7)	54.2(11.4)
	Pi-HAR	14.2(2.4)	24.9(6.6)	23.8(6.6)	23.4(5.5)	20.0(8.4)	42.3(12.6)	51.5(13.0)
	PL-HAR	13.7(3.0)	24.5(6.7)	24.0(6.8)	23.5(5.1)	22.0(9.2)	42.1(13.7)	51.5(14.2)
	VAT-HAR	13.2(2.9)	24.8(7.0)	24.2(6.0)	24.6(8.4)	20.6(7.9)	40.4(13.2)	51.4(11.7)
	VATENT-HAR	14.8(4.5)	25.5(4.3)	28.0(7.5)	25.4(5.9)	28.5(5.1)	53.5(10.9)	53.4(8.0)
	MixHAR(w/o Cal)	15.5(4.1)	28.3(6.2)	32.9(5.8)	28.2(7.1)	30.0(9.2)	51.8(15.1)	54.3(12.9)
	MixHAR	**17.0(3.8)**	**29.2(6.0)**	**33.9(7.6)**	**34.5(6.8)**	**36.3(7.0)**	**55.3(10.6)**	**56.4(7.1)**
Opportunity	Supervised	18.2(3.5)	23.0(4.0)	28.2(3.3)	30.8(2.7)	32.8(6.0)	40.3(2.7)	41.2(3.4)
	MT-HAR	19.1(2.9)	23.0(4.0)	28.2(2.6)	30.5(4.3)	32.3(5.1)	38.0(2.0)	41.0(2.1)
	Pi-HAR	16.1(1.6)	24.4(3.1)	27.4(2.8)	30.1(2.7)	31.0(4.8)	39.5(2.6)	40.0(2.2)
	PL-HAR	15.0(2.2)	24.0(3.1)	27.3(3.0)	30.8(4.2)	32.9(5.7)	40.2(3.6)	41.5(4.7)
	VAT-HAR	17.0(2.3)	24.5(3.4)	27.1(2.7)	29.7(4.4)	30.9(5.5)	39.0(2.7)	40.5(3.0)
	VATENT-HAR	10.3(1.4)	17.2(3.4)	26.9(5.6)	30.4(2.6)	32.5(5.6)	39.1(0.8)	39.4(2.8)
	MixHAR(w/o Cal)	19.3(3.0)	23.3(3.5)	29.4(2.8)	31.8(3.3)	34.3(4.3)	40.8(3.8)	41.3(4.1)
	MixHAR	**20.1(3.5)**	**25.4(2.5)**	**31.5(1.7)**	**32.0(3.1)**	**35.5(3.9)**	**40.9(3.0)**	**43.4(3.9)**

dataset, data collected from only 4 users in a constrained environment may not cause too many intra-activities issues, yet the main challenge may be the discrimination of the different non-repetitive activities. Although such performance of these conventional Deep-Semi-HAR was not always outstanding, results still showed their feasibility on making use of unlabelled activities, and they can be further improved in future work.

- Comparing with the supervised baseline on ALL different datasets with different labelled/unlabelled settings, our proposed MixHAR obtained significant mean F1 improvement and satisfactory standard deviation reduced, which suggests the advance of our method on taking advantage of unlabelled data in a deep semi-supervised fashion. We can see 0.6%-3.3% F_m improvement on Opportunity dataset, 0.8%-11.1% F_m improvement on PAMAP2 dataset, 5.3%-12.8% F_m improvement on mHealth dataset, 3.8%-10.6% F_m improvement on mHealth+ dataset, and 3.9%-9.7% F_m improvement on DSADS dataset. The improvement on the Opportunity dataset was not large as others, the possible reason may be that the activities in Opportunity are non-repetitive and extremely imbalanced.

- Comparing with the all the Deep-Semi-HAR baselines [111, 204, 114, 150], MixHAR still outperformed them with an obvious gap of F_m. Standard deviation indicated the model performance fluctuating on different subjects. Since these five approaches mainly focused on different consistency regularization for intra-activity variability, in some cases, they obtained slightly better standard deviation than ours.

- ***Effect of mixing calibration*** Table.6.1 also shows the effect of the proposed mixing calibration. As we can see, the main performance boosting in our MixHAR came from the labelled/unlabelled data mixing (MixHAR w/o Cal in Table.6.1), which is reasonable as our goal is to take advantage of unlabelled data. With help of the calibration for the activity-intrusion problem in feature space, we can see MixHAR obtained the satisfactory improvement of 0.1%-5.7% on different datasets under different settings, which suggests the importance of the calibration for labelled and unlabelled data interpolation.

- ***Effect of NULL-Class*** In the real-world scenario of ubiquitous HAR, most collected activities are irrelevant to the HAR system (NULL-Class) and there are only a few activities of interest [15]. There are two factors that the NULL-Class may affect recognition performance (e.g., on Opportunity and mHealth+ dataset). On the one hand, it is extremely imbalanced (e.g., more than 70% of the entire dataset) causing biased learning of the deep HAR model. Moreover, when the number of labelled data is very small,

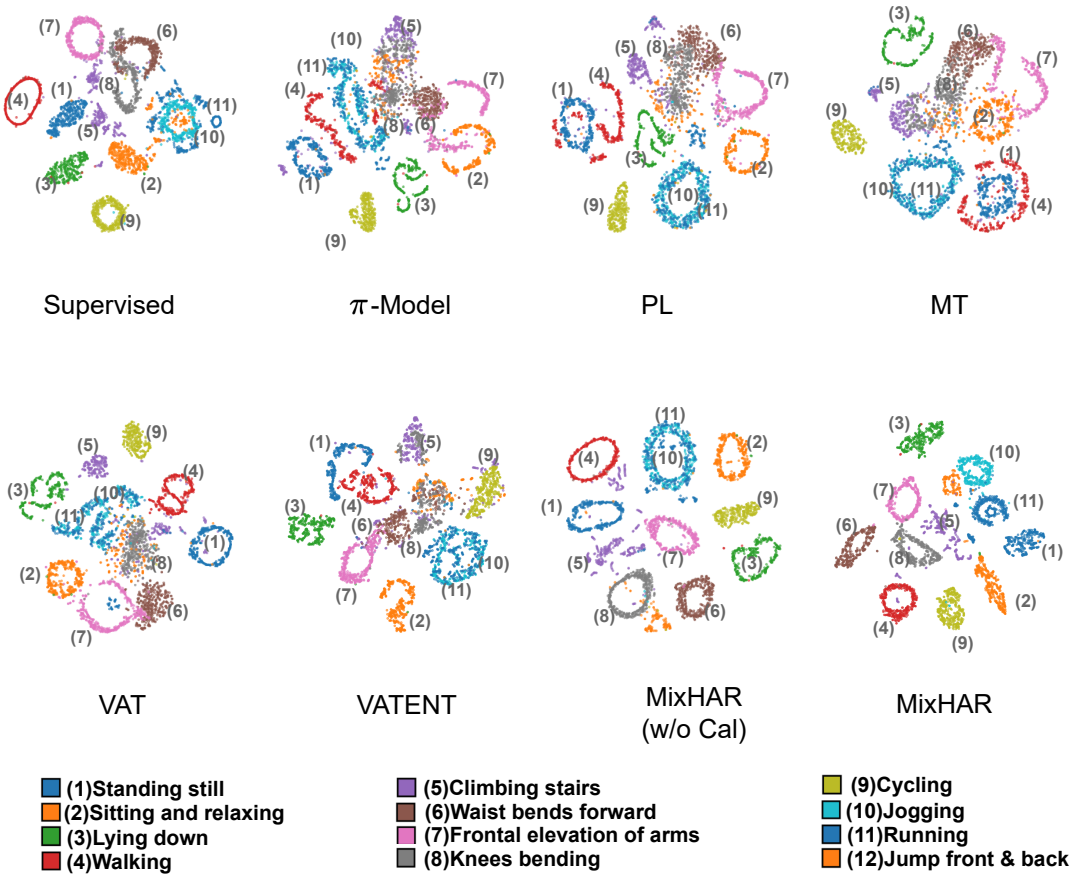


Fig. 6.3 Feature diagram on mHealth dataset based on 1% labelled data setting.

it is possible that most of the labelled data is NULL-Class activities, yet the activities of interest are the real targets and may very likely be under-represented/ignored during the training , which resulted in low F_m. On the other hand, the NULL-Class has similar patterns as activities of interest (but is unrelated to application) [15], which may lead to ambiguity during the model training. The opportunity dataset contains lots of NULL-Class activities and we can see that the overall mean F1 score of different methods on it were low, while the effect of our method (MixHAR) are obvious. The activities of interest in mHealth are mostly balanced, we used mHealth and its version that with NULL-Class (mHealth+) to reveal the effect of NULL-Class. In Table.6.1, the mean F1 score of different methods on mHealth+ (with NULL-Class) was largely degraded comparing with the corresponding setting on the mHealth (without NULL-Class). However, our MixHAR still obtained impressive performance than the supervised baseline and other Deep-Semi-HAR.

6.5.2 Feature Embedding Overview

Fig.6.3 is the visualisation of the learned feature embedding diagram of different Deep-Semi-HAR methods (including proposed MixHAR) based on mHealth dataset.

Model's Generalisation A robust deep HAR model should be able to handle the inter/intra-activity variability, as reflected in the feature diagram, the same classes of activities should be clustered together as tight as possible and different classes of activities should be separated apart distinctly. The activities in the mHealth dataset are repetitive (e.g., walking, running), as shown in Fig.6.3, more than half categories seem easy to be discriminated by the supervised baseline. There are obvious cons and pros of π-Model, PL and MT approaches. They all performed better on recognising the cycling by clear intra-activity distance reduced. However, the injected variability from unlabelled data may be too noisy to be handled and may misguide the representation learning. VAT and VATENT obtained better performance than the supervised baseline by better handling the intra-activity variability on walking, cycling and climbing stairs. Most importantly, the robustness of the model trained by MixHAR was considerably improved with enlarging the inter-activity distance and reducing the intra-activity distance on all different activities, which suggests the effectiveness of MixHAR to take advantage of unlabelled data and improve the model's generalisation.

Effect of Mixing Calibration Mixing labelled/unlabelled data injects the large variability for robust model training with keeping discriminative classification information. Since the labelled data are limited and mixing substantial unlabelled data may challenge the model's capability of handling diverse information. Specifically, when activity-intrusion happens, the conflicts between the mixed activities and original activities may inhibit the model's discrimination on different activities, which is still weak on different activities with similar characteristic such as knees bending and jump front & back, running and jogging (MixHAR w/o Cal in Fig.6.3). Our proposed mixing calibration aims to learn the correlation of mixed samples on feature space and enhance the model's discrimination on injected variability. With further mixing calibration (MixHAR in Fig.6.3), it was very clear that all the different activities (inter-activity) were discriminated obviously by MixHAR with intra-activity distance reduced and inter-activity distance enlarged.

6.5.3 Improvement of Minority-Activity-Classes

Fig.6.4 shows the benefits of using MixHAR for each class in the practical case that the data are imbalanced (typically with more than 70% NULL-Class). For both Opportunity and mHealth+ data, using unlabelled data by MixHAR (red/purple in Fig.6.4) obtained better generalisation on most minority classes with keeping the performance on majority classes.

For the Opportunity dataset, although the model's generalisation was improved for most activities by MixHAR, the model's performance still mainly came from the imbalanced NULL-Class activities. The reason can be that in the Opportunity dataset, most of the NULL-Class activities is walking or other simple transitional activities [19], and other activities of interests are relatively complex (i.e., non-repetitive) [19]. For the mHealth+ dataset, we can observe the obvious the improvement of MixHAR. It is interesting that when the labelled data are limited, leveraging extra unlabelled data may still not be effective for the activities that are relatively simple like standing still. These activities have the characteristics that may widely exist in lots of other activities, which may be mis-recognised.

We also compared our MixHAR with the standard random oversampling (orange in Fig.6.4) and undersampling [14, 21] (green in Fig.6.4) method. As we can see, oversampling and undersampling is feasible to improve the minority class however our method still outperformed them on nearly all the classes of activities. Moreover, oversampling/undersampling methods largely degraded the model performance on the head class (NULL-Class) while our method also improved the performance on it. This result suggests that our deep semi-supervised method MixHAR may be able to handle the largely imbalanced activities that can be a valuable research direction in real-world ubiquitous HAR.

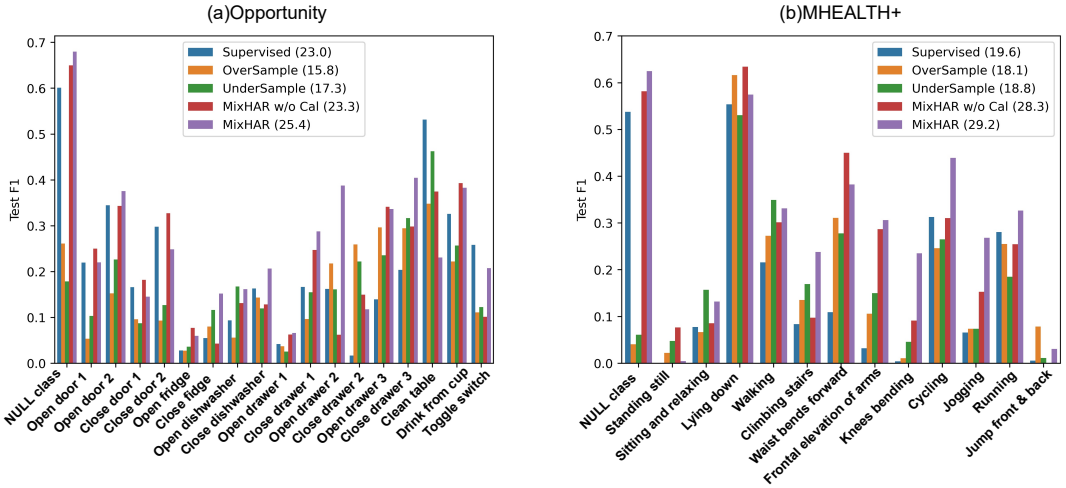


Fig. 6.4 Class-wise recognition results on the data imbalanced dataset Opportunity and mHealth+ based on 1% labelled data setting. The mean F1 score is placed in brackets.

6.5.4 Taking advantage of unlabelled data

In recent years, taking advantage of unlabelled data has attracted growing research attention in deep HAR [7, 181, 71, 160]. Specifically, in addition to semi-supervised HAR,

Table 6.2 Comparison of the approaches that take advantage of unlabelled data with 1% labelled data, which is based on our deep semi-supervised settings/pipelines.

methods	Opportunity	PAMAP2	mHealth	mHealth+	DSADS
AAE-HAR[7]	10.6(**1.1**)	16.3(**5.3**)	26.9(5.8)	13.1(**0.7**)	18.4(**2.3**)
MTL-Self-HAR[181]	13.5(1.2)	18.5(6.57)	39.1(**5.2**)	17.9(4.9)	20.1(3.1)
CL-HAR[250]	11.4(1.5)	17.3(6.1)	42.9(10.8)	24.4(4.7)	39.4(7.6)
Mask-Self-HAR[71]	15.9(1.76)	32.8(9.7)	41.3(8.6)	25.5(6.8)	51.4(6.1)
Selfhar[203]	20.9(4.8)	40.3(12.4)	44.2(7.9)	24.6(5.4)	53.3(1.4)
MixHAR	**25.4**(2.5)	38.7(6.7)	**46.3**(10.3)	**29.2**(6.0)	**59.5**(6.3)

self-supervised learning is also obtained rapidly growing attention. In order to evaluate the advance of our MixHAR, here we also compared MixHAR against the recent state-of-the-art methods taking advantage of unlabelled data, which includes but is not limited to semi-supervised HAR (i.e., self-supervised HAR). We developed the five recent existing works on five wearable-based HAR datasets. Adversarial Auto-Encoder (AAE-HAR) focued on learning the activity representations by an encoder-decoder network with two discriminators handling the style information [7]. A multi-task self-supervised learning framework was proposed to pre-train an activity representation learner by discriminating the various transformations that applied on input signals [181], and the representation learner was used on downstream activity classification task (MTL-HAR). Motivated by ladder net [175], Zeng et al. also proposed a CNN encoder-decoder network with noisy injecting and cleaning to take advantage of unlabelled data [250]. Also, Haresamudram et al. proposed a self-supervised method to pre-train an activity representation learner based on reconstructing unlabelled data in a mask-then-reconstruct manner [71], and then the trained learner was utilised on downstream activity classification task (Mask-Self-HAR). Most recently, Tang et al. proposed a Selfhar framework which is based on not only the existing self-supervised technologies but also semi-supervised training to leverage the unlabeled human activities [203]. As shown in Table.6.2, we can see that our MixHAR outperformed all these methods on the mean F1 score with the slightly higher standard deviation. In AAE, it is hard to judge what information exactly the discriminators learned, it seems that when the number of labelled data is extremely small (e.g., 1% of entire training data) the discriminator may not be fitted and negatively affect the final recognition performance. For CL-HAR net, the encoder-decoder based on the noisy signals may lose the information of original signals, and the reconstruction process was always biased. Analogously, the MTL-HAR, which built the

self-supervised framework based on general data augmentation, may suffer from designing the HAR-irrelevant pre-training tasks. As for Mask-Self-HAR, reconstructing the original signals helped the model to learn the relatively accurate patterns of different activities, which obtained satisfactory performance. Also, most recent work, i.e., Selfhar, obtained clear improvement compared with previous methods. However, the the weak HAR-specialise information during the pre-training may still result in less robust features and not strong enough performance on relatively complex data for these self-supervised based methods. Moreover, when with NULL-Class activities, the activities of interests may be largely under-represented during the pre-training stage. Hence, these self-supervised based methods' performance is still lower than ours when with non-repetitive activities (e.g., Opportunity dataset) or NULL-Class activities, which suggests the more robust and reliable of our approach.

6.6 Conlusion

Based on the recent success of semi-supervised techniques in deep learning, we developed and evaluated the five conventional/popular deep semi-supervised approach on ubiquitous HAR (Deep-Semi-HAR) with discussing cons and pros of them. Most importantly, we proposed a novel deep semi-supervised approach named MixHAR to take advantage of unlabelled activities MixHAR leverages the linear interpolation to mix the labelled and unlabelled data simultaneously with corresponding mixing calibration, which not only ensures the model's discrimination of different activities but also improve the model's capability on handling diverse variability. Comparing with both conventional Deep-Semi-HAR methods and recent works that also make use of unlabelled data, results show that MixHAR is more effective and powerful for sensor-based human activity recognition.

Chapter 7

Conclusion

This thesis has made significant strides in advancing the field of deep learning, particularly in the application of attention mechanisms across various domains. Drawing inspiration from the human perceptual system, which effectively manages vast amounts of information through selective attention, this work has successfully implemented analogous strategies in deep artificial neural networks. These implementations have shown that models can be trained to focus on the most pertinent parts of signals, such as in image recognition and time series data modeling, thereby improving performance and computational efficiency.

Chapters 3 and 4 dealt with the application of attention mechanisms in crowd counting, demonstrating how dual attention mechanisms can enhance representation learning. The incorporation of second-order and first-order attention into a multi-stream architecture marks a notable advancement in this area. Chapter 4 further explored the use of attention in semi-supervised learning contexts, emphasizing the potential of unlabelled data in deep model training. This approach not only harnesses rich information but also reduces the cost and labor associated with data annotation. The unified attention model presented in Chapter 5 is a groundbreaking contribution, combining convolution operations and transformer layers to create a hybrid model adept at handling multi-level features adaptively. This model represents a significant leap in visual feature modeling, capable of dynamically enhancing positive local patches and identifying local co-occurrences. Chapter 6 extended the application of attention mechanisms to time series data, particularly in the context of human activity recognition using wearable sensors. This novel approach demonstrates the potential of attention mechanisms in identifying and emphasizing salient features in complex datasets, thereby enhancing predictive accuracy in dynamic environments. This chapter further proved the proposed attention mechanism is effective not only in the vision domain but also in the time series domain.

Despite these achievements, several challenges and avenues for future work remain. The balance between accuracy and computational efficiency continues to be a primary concern, especially in the context of non-local attention mechanisms. Addressing scalability and computational cost issues in high-resolution images or lengthy sequences remains a critical challenge. Moreover, enhancing the generalizability of attention models across different tasks and domains is an area that warrants further exploration. The development of flexible attention mechanisms that can adapt to varied inputs and contexts, such as large-scale scene understanding or multimodal alignment, presents an exciting frontier for research. Future work should focus on designing attention models that are not only effective in specific applications but also exhibit versatility and adaptability across various tasks. Developing methods to dynamically alter the focus of models, depending on the complexity and nature of datasets, is imperative. Moreover, further exploration into the integration of attention mechanisms in semi-supervised and unsupervised learning settings could pave the way for more efficient and cost-effective models. Lastly, extending these concepts to other data modalities, beyond visual and time-series data, could open new horizons in the field of deep learning and artificial intelligence.

In conclusion, this thesis contributes significantly to the body of knowledge in deep learning and attention mechanisms, presenting innovative solutions and paving the way for future advancements in the field. The proposed models and mechanisms not only enhance current understanding and capabilities but also set the stage for continued innovation and exploration in this rapidly evolving domain.

References

[1] Aleissaee, A. A., Kumar, A., Anwer, R. M., Khan, S., Cholakkal, H., Xia, G.-S., et al. (2022). Transformers in remote sensing: A survey. *arXiv preprint arXiv:2209.01206*.

[2] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

[3] Azad, R., Asadi-Aghbolaghi, M., Fathy, M., and Escalera, S. (2019). Bi-directional convlstm u-net with densley connected convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0.

[4] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[5] Bai, L., Yao, L., Wang, X., Kanhere, S. S., Guo, B., and Yu, Z. (2020a). Adversarial multi-view networks for activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(2):1–22.

[6] Bai, S., He, Z., Qiao, Y., Hu, H., Wu, W., and Yan, J. (2020b). Adaptive dilated network with self-correction supervision for counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4594–4603.

[7] Balabka, D. (2019). Semi-supervised learning for human activity recognition using adversarial autoencoders. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, pages 685–688.

[8] Banos, O., Garcia, R., Holgado-Terriza, J. A., Damas, M., Pomares, H., Rojas, I., Saez, A., and Villalonga, C. (2014). mhealthdroid: a novel framework for agile development of mobile health applications. In *International workshop on ambient assisted living*, pages 91–98. Springer.

[9] Bao, L. and Intille, S. S. (2004). Activity recognition from user-annotated acceleration data. In *International conference on pervasive computing*, pages 1–17. Springer.

[10] Barshan, B. and Yüksek, M. C. (2014). Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units. *The Computer Journal*, 57(11):1649–1667.

[11] Bello, I., Zoph, B., Vaswani, A., Shlens, J., and Le, Q. V. (2019). Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3286–3295.

[12] Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5049–5059.

[13] Bhattacharya, S., Nurmi, P., Hammerla, N., and Plötz, T. (2014). Using unlabeled data in a sparse-coding framework for human activity recognition. *Pervasive and Mobile Computing*, 15:242–262.

[14] Buda, M., Maki, A., and Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259.

[15] Bulling, A., Blanke, U., and Schiele, B. (2014). A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)*, 46(3):1–33.

[16] Cao, X., Wang, Z., Zhao, Y., and Su, F. (2018). Scale aggregation network for accurate and efficient crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750.

[17] Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168.

[18] Chapelle, O., Scholkopf, B., and Zien, A. (2009). Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542.

[19] Chavarriaga, R., Sagha, H., Calatroni, A., Digumarti, S. T., Tröster, G., Millán, J. d. R., and Roggen, D. (2013). The opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters*, 34(15):2033–2042.

[20] Chavdarova, T. and Fleuret, F. (2018). Sgan: An alternative training of generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9407–9415.

[21] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

[22] Chefer, H., Gur, S., and Wolf, L. (2021). Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406.

[23] Chen, C.-F., Panda, R., and Fan, Q. (2021). Regionvit: Regional-to-local attention for vision transformers.

[24] Chen, J., Yang, Z., and Yang, D. (2020a). Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *Annual Meeting of the Association for Computational Linguistics*.

[25] Chen, K., Yao, L., Zhang, D., Wang, X., Chang, X., and Nie, F. (2019a). A semisupervised recurrent convolutional attention model for human activity recognition. *IEEE transactions on neural networks and learning systems*, 31(5):1747–1756.

[26] Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., and Liu, Y. (2020b). Deep learning for sensor-based human activity recognition: overview, challenges and opportunities. *arXiv preprint arXiv:2001.07416*.

[27] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018a). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818.

[28] Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., and Wang, Z. (2019b). Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8351–8361.

[29] Chen, X., Bin, Y., Sang, N., and Gao, C. (2019c). Scale pyramid network for crowd counting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1941–1950. IEEE.

[30] Chen, X., Fu, C., Zhao, Y., Zheng, F., Song, J., Ji, R., and Yang, Y. (2020c). Salience-guided cascaded suppression network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3300–3310.

[31] Chen, Z., Wang, S., Hou, X., Shao, L., and Dhabi, A. (2018b). Recurrent transformer network for remote sensing scene categorisation. In *British Machine Vision Conference (BMVC)*, page 266.

[32] Cheng, C.-C., Qiu, M.-X., Chiang, C.-K., and Lai, S.-H. (2023). Rest: A reconfigurable spatial-temporal graph model for multi-camera multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10051–10060.

[33] Cheng, Z.-Q., Li, J.-X., Dai, Q., Wu, X., and Hauptmann, A. G. (2019a). Learning spatial awareness to improve crowd counting. In *The IEEE International Conference on Computer Vision (ICCV)*.

[34] Cheng, Z.-Q., Li, J.-X., Dai, Q., Wu, X., He, J.-Y., and Hauptmann, A. G. (2019b). Improving the learning of multi-column convolutional neural network for crowd counting. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1897–1906. ACM.

[35] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

[36] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

[37] Contributors, M. (2020). Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark.

[38] Corbetta, M. and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201–215.

[39] Cordonnier, J.-B., Loukas, A., and Jaggi, M. (2020). On the relationship between self-attention and convolutional layers. (CONF).

[40] Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703.

[41] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., and Wei, Y. (2017a). Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773.

[42] Dai, X., Yue-Hei Ng, J., and Davis, L. S. (2017b). Fason: First and second order information fusion network for texture recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[43] Dai, Z., Liu, H., Le, Q., and Tan, M. (2021). Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34.

[44] Dang, L. M., Min, K., Wang, H., Piran, M. J., Lee, C. H., and Moon, H. (2020). Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition*, 108:107561.

[45] Dawn, D. D. and Shaikh, S. H. (2016). A comprehensive survey of human action recognition with spatio-temporal interest point (stip) detector. *The Visual Computer*, 32(3):289–306.

[46] Deng, J., Guo, J., and Zafeiriou, S. (2019). Single-stage joint face detection and alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.

[47] Ding, X., He, F., Lin, Z., Wang, Y., Guo, H., and Huang, Y. (2020). Crowd density estimation using fusion of multi-layer features. *IEEE Transactions on Intelligent Transportation Systems*, 22(8):4776–4787.

[48] Dong, L., Zhang, H., Ji, Y., and Ding, Y. (2020). Crowd counting by using multi-level density-based spatial information: A multi-scale cnn framework. *Information Sciences*, 528:79–91.

[49] Dong, L., Zhang, H., Ma, J., Xu, X., Yang, Y., and Wu, Q. J. (2022a). Clrnet: a cross locality relation network for crowd counting in videos. *IEEE Transactions on Neural Networks and Learning Systems*.

[50] Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., and Guo, B. (2022b). Cswin transformer: A general vision transformer backbone with cross-shaped windows. pages 12124–12134.

[51] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale.

[52] Duan, H., Wang, S., and Guan, Y. (2020). Sofa-net: Second-order and first-order attention network for crowd counting. *British Machine Vision Conference (BMVC)*.

[53] Elsken, T., Metzen, J. H., and Hutter, F. (2019). Neural architecture search: A survey. *The Journal of Machine Learning Research*, 20(1):1997–2017.

[54] Fu, J., Liu, J., Wang, Y., Li, Y., Bao, Y., Tang, J., and Lu, H. (2019). Adaptive context network for scene parsing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6748–6757.

[55] Gao, J., Wang, Q., and Yuan, Y. (2019a). Scar: Spatial-/channel-wise attention regression networks for crowd counting. *Neurocomputing*, 363:1–8.

[56] Gao, Y., Long, Y., Guan, Y., Basu, A., Baggaley, J., and Ploetz, T. (2019b). Towards reliable, automated general movement assessment for perinatal stroke screening in infants using wearable accelerometers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(1).

[57] Gao, Z., Xie, J., Wang, Q., and Li, P. (2019c). Global second-order pooling convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033.

[58] Girdhar, R., Singh, M., Ravi, N., van der Maaten, L., Joulin, A., and Misra, I. (2022). Omnivore: A single model for many visual modalities. pages 16102–16112.

[59] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

[60] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

[61] Grandvalet, Y. and Bengio, Y. (2004). Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17.

[62] Grandvalet, Y., Bengio, Y., et al. (2005). Semi-supervised learning by entropy minimization. In *CAP*, pages 281–296.

[63] Guan, D., Yuan, W., Lee, Y.-K., Gavrilov, A., and Lee, S. (2007). Activity recognition based on semi-supervised learning. In *13th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA 2007)*, pages 469–475. IEEE.

[64] Guan, Y., Li, C.-T., and Hu, Y. (2012). Robust clothing-invariant gait recognition. In *2012 eighth international conference on intelligent information hiding and multimedia signal processing*, pages 321–324. IEEE.

[65] Guan, Y. and Plötz, T. (2017). Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):1–28.

[66] Guo, H., Mao, Y., and Zhang, R. (2019). Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3714–3722.

[67] Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., and Xu, C. (2022a). Cmt: Convolutional neural networks meet vision transformers. pages 12175–12185.

[68] Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., Zhang, S.-H., Martin, R. R., Cheng, M.-M., and Hu, S.-M. (2022b). Attention mechanisms in computer vision: A survey. *Computational visual media*, 8(3):331–368.

[69] Hammerla, N. Y., Halloran, S., and Plötz, T. (2016). Deep, convolutional, and recurrent models for human activity recognition using wearables. *International Joint Conference on Artificial Intelligence (IJCAI)*.

[70] Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., and Wang, Y. (2021). Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919.

[71] Haresamudram, H., Beedu, A., Agrawal, V., Grady, P. L., Essa, I., Hoffman, J., and Plötz, T. (2020). Masked reconstruction based self-supervision for human activity recognition. In *Proceedings of the 2020 International Symposium on Wearable Computers*, pages 45–49.

[72] Hassanin, M., Anwar, S., Radwan, I., Khan, F. S., and Mian, A. (2022). Visual attention methods in deep learning: An in-depth survey. *arXiv preprint arXiv:2204.07756*.

[73] Hayhoe, M. and Ballard, D. (2005). Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4):188–194.

[74] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[75] He, S., Luo, H., Wang, P., Wang, F., Li, H., and Jiang, W. (2021). Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15013–15022.

[76] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

[77] Herath, S., Harandi, M., and Porikli, F. (2017). Going deeper into action recognition: A survey. *Image and vision computing*, 60:4–21.

[78] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

[79] Hossain, M. Z., Sohel, F., Shiratuddin, M. F., and Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36.

[80] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

[81] Hu, H., Zhang, Z., Xie, Z., and Lin, S. (2019). Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3464–3473.

[82] Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

[83] Hu, Y., Jiang, X., Liu, X., Zhang, B., Han, J., Cao, X., and Doermann, D. (2020). Nas-count: Counting-by-density with neural architecture search. *European conference on computer vision*.

[84] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

[85] Huang, Z., Ben, Y., Luo, G., Cheng, P., Yu, G., and Fu, B. (2021). Shuffle transformer: Rethinking spatial shuffle for vision transformer. *arXiv preprint arXiv:2106.03650*.

[86] Huynh, T. and Schiele, B. (2005). Analyzing features for activity recognition. In *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies*, pages 159–163.

[87] Idrees, H., Saleemi, I., Seibert, C., and Shah, M. (2013). Multi-source multi-scale counting in extremely dense crowd images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[88] Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., and Shah, M. (2018). Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–546.

[89] Ionescu, C., Vantzos, O., and Sminchisescu, C. (2015). Matrix backpropagation for deep networks with structured layers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2965–2973.

[90] Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259.

[91] Ji, S., Xu, W., Yang, M., and Yu, K. (2012). 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231.

[92] Jia, X., Jing, X.-Y., Zhu, X., Chen, S., Du, B., Cai, Z., He, Z., and Yue, D. (2020). Semi-supervised multi-view deep discriminant representation learning. *IEEE transactions on pattern analysis and machine intelligence*, 43(7):2496–2509.

[93] Jiang, X., Zhang, L., Xu, M., Zhang, T., Lv, P., Zhou, B., Yang, X., and Pang, Y. (2020). Attention scaling for crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4706–4715.

[94] Jin, X., Lan, C., Zeng, W., Wei, G., and Chen, Z. (2020). Semantics-aligned representation learning for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11173–11180.

[95] Kanwisher, N. and Wojciulik, E. (2000). Visual attention: insights from brain imaging. *Nature reviews neuroscience*, 1(2):91–100.

[96] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. (2021). Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*.

[97] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. (2022). Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41.

[98] Kim, W. and Lee, Y. (2019). Learning dynamics of attention: Human prior for interpretable machine reasoning. *Advances in Neural Information Processing Systems*, 32.

[99] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

[100] Kingma, D. P., Mohamed, S., Jimenez Rezende, D., and Welling, M. (2014). Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27.

[101] Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

[102] Kirillov, A., Girshick, R., He, K., and Dollár, P. (2019). Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6399–6408.

[103] Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. (2020). Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer.

[104] Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, 39:261–283.

[105] Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

[106] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

[107] Kwapisz, J. R., Weiss, G. M., and Moore, S. A. (2011). Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82.

[108] Kwon, H., Abowd, G. D., and Plötz, T. (2019). Handling annotation uncertainty in human activity recognition. In *Proceedings of the 23rd International Symposium on Wearable Computers*, pages 109–117.

[109] Kwon, H., Tong, C., Haresamudram, H., Gao, Y., Abowd, G. D., Lane, N. D., and Ploetz, T. (2020). Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*.

[110] Lahoud, J., Cao, J., Khan, F. S., Cholakkal, H., Anwer, R. M., Khan, S., and Yang, M.-H. (2022). 3d vision with transformers: A survey. *arXiv preprint arXiv:2208.04309*.

[111] Laine, S. and Aila, T. (2016). Temporal ensembling for semi-supervised learning. *International Conference on Learning Representations*.

[112] Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8595–8598. IEEE.

[113] Lee, D.-H. (2013a). Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks.

[114] Lee, D.-H. (2013b). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3.

[115] Lempitsky, V. and Zisserman, A. (2010). Learning to count objects in images. In *Advances in neural information processing systems*, pages 1324–1332.

[116] Li, J., Socher, R., and Hoi, S. C. (2020). Dividemix: Learning with noisy labels as semi-supervised learning. *International Conference on Learning Representations*.

[117] Li, P., Xie, J., Wang, Q., and Zuo, W. (2017a). Is second-order information helpful for large-scale visual recognition? In *The IEEE International Conference on Computer Vision (ICCV)*.

[118] Li, P., Xie, J., Wang, Q., and Zuo, W. (2017b). Is second-order information helpful for large-scale visual recognition? In *Proceedings of the IEEE international conference on computer vision*, pages 2070–2078.

[119] Li, W., Chen, H., Guo, J., Zhang, Z., and Wang, Y. (2022). Brain-inspired multilayer perceptron with spiking neurons. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 783–793.

[120] Li, Y., Zhang, K., Cao, J., Timofte, R., and Van Gool, L. (2021a). Localvit: Bringing locality to vision transformers.

[121] Li, Y., Zhang, X., and Chen, D. (2018a). Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[122] Li, Y., Zhang, X., and Chen, D. (2018b). Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100.

[123] Li, Z., Liu, F., Yang, W., Peng, S., and Zhou, J. (2021b). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*.

[124] Lin, H., Ma, Z., Ji, R., Wang, Y., and Hong, X. (2022). Boosting crowd counting via multifaceted attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19628–19637.

[125] Liu, C., Weng, X., and Mu, Y. (2019a). Recurrent attentive zooming for joint crowd counting and precise localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[126] Liu, L., Cao, Z., Lu, H., Xiong, H., and Shen, C. (2020a). Nssnet: Scale-aware object counting with non-scale suppression. *IEEE Transactions on Intelligent Transportation Systems*.

[127] Liu, L., Lu, H., Zou, H., Xiong, H., Cao, Z., and Shen, C. (2020b). Weighing counts: Sequential crowd counting by reinforcement learning. In *European Conference on Computer Vision*, pages 164–181. Springer.

[128] Liu, L., Qiu, Z., Li, G., Liu, S., Ouyang, W., and Lin, L. (2019b). Crowd counting with deep structured scale integration network. In *The IEEE International Conference on Computer Vision (ICCV)*.

[129] Liu, L., Qiu, Z., Li, G., Liu, S., Ouyang, W., and Lin, L. (2019c). Crowd counting with deep structured scale integration network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1774–1783.

[130] Liu, L., Wang, H., Li, G., Ouyang, W., and Lin, L. (2018a). Crowd counting using deep recurrent spatial-aware network. *International Joint Conference on Artificial Intelligence (IJCAI)*.

[131] Liu, W., Salzmann, M., and Fua, P. (2019d). Context-aware crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5099–5108.

[132] Liu, X., Van De Weijer, J., and Bagdanov, A. D. (2018b). Leveraging unlabeled data for crowd counting by learning to rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7661–7669.

[133] Liu, X., Yang, J., Ding, W., Wang, T., Wang, Z., and Xiong, J. (2020c). Adaptive mixture regression network with local counting map for crowd counting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 241–257. Springer.

[134] Liu, Y., Liu, L., Wang, P., Zhang, P., and Lei, Y. (2020d). Semi-supervised crowd counting via self-training on surrogate tasks. *Proceedings of the European Conference on Computer Vision (ECCV)*.

[135] Liu, Y., Wang, Z., Shi, M., Satoh, S., Zhao, Q., and Yang, H. (2020e). Towards unsupervised crowd counting via regression-detection bi-knowledge transfer. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 129–137.

[136] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. pages 10012–10022.

[137] Long, Y., Liu, L., Shen, F., Shao, L., and Li, X. (2017). Zero-shot learning using synthesised unseen visual data with diffusion regularisation. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2498–2512.

[138] Loshchilov, I. and Hutter, F. (2018). Decoupled weight decay regularization.

[139] Luo, A., Yang, F., Li, X., Nie, D., Jiao, Z., Zhou, S., and Cheng, H. (2020). Hybrid graph neural networks for crowd counting. *The Association for the Advancement of Artificial Intelligence (AAAI)*.

[140] Luo, W., Li, Y., Urtasun, R., and Zemel, R. (2016). Understanding the effective receptive field in deep convolutional neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4905–4913.

[141] Lv, Y., Li, M., He, Y., Li, S., He, Z., and Yang, A. (2023). Anchor-intermediate detector: Decoupling and coupling bounding boxes for accurate object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6275–6284.

[142] Ma, Y. and Ghasemzadeh, H. (2019). Labelforest: non-parametric semi-supervised learning for activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4520–4527.

[143] Ma, Z., Wei, X., Hong, X., and Gong, Y. (2019). Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6142–6151.

[144] Mehta, S. and Rastegari, M. (2021). Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer.

[145] Meng, Y., Zhang, H., Zhao, Y., Yang, X., Qian, X., Huang, X., and Zheng, Y. (2021). Spatial uncertainty-aware semi-supervised crowd counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15549–15559.

[146] Miao, J., Wu, Y., Liu, P., Ding, Y., and Yang, Y. (2019). Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 542–551.

[147] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., and Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542.

[148] Mishra, N., Schreiber, R., Stanton, I., and Tarjan, R. E. (2007). Clustering social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 56–67. Springer.

[149] Mittal, S., Tatarchenko, M., and Brox, T. (2019). Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[150] Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

[151] Mnih, V., Heess, N., Graves, A., et al. (2014). Recurrent models of visual attention. *Advances in neural information processing systems*, 27.

[152] Morshed, M. B., Saha, K., Li, R., D'Mello, S. K., De Choudhury, M., Abowd, G. D., and Plötz, T. (2019). Prediction of mood instability with passive sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–21.

[153] Mott, A., Zoran, D., Chrzanowski, M., Wierstra, D., and Jimenez Rezende, D. (2019). Towards interpretable reinforcement learning using attention augmented agents. *Advances in neural information processing systems*, 32.

[154] Ng, A. Y. (2004). Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78.

[155] Nilsback, M.-E. and Zisserman, A. (2008). Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE.

[156] Noble, D. (2002). The rise of computational biology. *Nature Reviews Molecular Cell Biology*, 3(6):459–463.

[157] Novikov, A., Podoprikhin, D., Osokin, A., and Vetrov, D. P. (2015). Tensorizing neural networks. *Advances in neural information processing systems*, 28.

[158] Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., and Goodfellow, I. (2018). Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31:3235–3246.

[159] Ordóñez, F. J. and Roggen, D. (2016). Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115.

[160] Ouali, Y., Hudelot, C., and Tami, M. (2020). An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*.

[161] Pan, F., Shin, I., Rameau, F., Lee, S., and Kweon, I. S. (2020). Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3764–3773.

[162] Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. (2012). Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE.

[163] Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., and Tran, D. (2018). Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR.

[164] Petersen, S. E. and Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual review of neuroscience*, 35:73–89.

[165] Pham, H., Dai, Z., Xie, Q., and Le, Q. V. (2021). Meta pseudo labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11557–11568.

[166] Pham, V.-Q., Kozakaya, T., Yamaguchi, O., and Okada, R. (2015). Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In *The IEEE International Conference on Computer Vision (ICCV)*.

[167] Piroonsup, N. and Sinthupinyo, S. (2018). Analysis of training data using clustering to improve semi-supervised self-training. *Knowledge-Based Systems*, 143:65–80.

[168] PlÖtz, T. (2021). Applying machine learning for sensor data analysis in interactive systems: Common pitfalls of pragmatic use and ways to avoid them. *ACM Computing Surveys (CSUR)*, 54(6):1–25.

[169] Plötz, T., Hammerla, N. Y., and Olivier, P. L. (2011). Feature learning for activity recognition in ubiquitous computing. In *Twenty-second international joint conference on artificial intelligence*.

[170] Posner, M. I. and Petersen, S. E. (1990). The attention system of the human brain. *Annual review of neuroscience*, 13(1):25–42.

[171] Qian, B., Su, J., Wen, Z., Jha, D. N., Li, Y., Guan, Y., Puthal, D., James, P., Yang, R., Zomaya, A. Y., et al. (2020). Orchestrating the development lifecycle of machine learning-based iot applications: A taxonomy and survey. *ACM Computing Surveys (CSUR)*.

[172] Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., and Dollár, P. (2020). Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436.

[173] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

[174] Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34.

[175] Rasmus, A., Valpola, H., Honkala, M., Berglund, M., and Raiko, T. (2015). Semi-supervised learning with ladder networks. *The Conference on Neural Information Processing Systems (NeurIPS)*.

[176] Ray, P., Reddy, S. S., and Banerjee, T. (2021). Various dimension reduction techniques for high dimensional data analysis: a review. *Artificial Intelligence Review*, 54:3473–3515.

[177] Reiss, A. and Stricker, D. (2012). Introducing a new benchmarked dataset for activity monitoring. In *2012 16th International Symposium on Wearable Computers*, pages 108–109. IEEE.

[178] Rensink, R. A. (2000). The dynamic representation of scenes. *Visual cognition*, 7(1-3):17–42.

[179] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

[180] Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. *Advances in neural information processing systems*, 30.

[181] Saeed, A., Ozcelebi, T., and Lukkien, J. (2019). Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2):1–30.

[182] Sam, D. B., Sajjan, N. N., Maurya, H., and Babu, R. V. (2019). Almost unsupervised learning for dense crowd counting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8868–8875.

[183] Sam, D. B., Surya, S., and Babu, R. V. (2017). Switching convolutional neural network for crowd counting. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4031–4039. IEEE.

[184] Sangam, T., Dave, I. R., Sultani, W., and Shah, M. (2022). Transvisdrone: Spatio-temporal transformer for vision-based drone-to-drone detection in aerial videos. *arXiv preprint arXiv:2210.08423*.

[185] Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

[186] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

[187] Shao, Z., Zhang, X., Ding, C., Wang, J., and Wang, J. (2023). Unified pre-training with pseudo texts for text-to-image person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11174–11184.

[188] Sheng, T. and Huber, M. (2020). Weakly supervised multi-task representation learning for human activity analysis using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(2):1–18.

[189] Shi, M., Yang, Z., Xu, C., and Chen, Q. (2019). Revisiting perspective information for efficient crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7279–7288.

[190] Si, C., Nie, X., Wang, W., Wang, L., Tan, T., and Feng, J. (2020). Adversarial self-supervised learning for semi-supervised 3d action recognition. *European conference on computer vision*.

[191] Simonyan, K. and Zisserman, A. (2014a). Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27:568–576.

[192] Simonyan, K. and Zisserman, A. (2014b). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*.

[193] Sindagi, V. A. and Patel, V. M. (2019). Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *The IEEE International Conference on Computer Vision (ICCV)*.

[194] Sindagi, V. A., Yasarla, R., Babu, D. S., Babu, R. V., and Patel, V. M. (2020). Learning to count in the crowd from limited labeled data. *Proceedings of the European Conference on Computer Vision (ECCV)*.

[195] Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *The Conference on Neural Information Processing Systems (NeurIPS)*.

[196] Song, Q., Wang, C., Jiang, Z., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., and Wu, Y. (2021). Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3365–3374.

[197] Srinivas, A., Lin, T.-Y., Parmar, N., Shlens, J., Abbeel, P., and Vaswani, A. (2021). Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16519–16529.

[198] Srivastava, N., Mansimov, E., and Salakhudinov, R. (2015). Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852.

[199] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. (2019). Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

[200] Sultana, M., Naseer, M., Khan, M. H., Khan, S., and Khan, F. S. (2022). Self-distilled vision transformer for domain generalization. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[201] Sun, G., Liu, Y., Probst, T., Paudel, D. P., Popovic, N., and Van Gool, L. (2021). Boosting crowd counting with transformers. *arXiv preprint arXiv:2105.10926*.

[202] Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR.

[203] Tang, C. I., Perez-Pozuelo, I., Spathis, D., Brage, S., Wareham, N., and Mascolo, C. (2021). Selfhar: Improving human activity recognition through self-training with unlabeled data. *arXiv preprint arXiv:2102.06073*.

[204] Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204.

[205] Tian, Y., Chu, X., and Wang, H. (2021). Cctrans: Simplifying and improving crowd counting with transformer. *arXiv preprint arXiv:2109.14483*.

[206] Tian, Y., Yu, X., Fan, B., Wu, F., Heijnen, H., and Balntas, V. (2019). Sosnet: Second order similarity regularization for local descriptor learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11016–11025.

[207] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR.

[208] Um, T. T., Pfister, F. M., Pichler, D., Endo, S., Lang, M., Hirche, S., Fietzek, U., and Kulić, D. (2017). Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 216–220.

[209] Van Engelen, J. E. and Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440.

[210] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[211] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

[212] Verma, V., Lamb, A., Kannala, J., Bengio, Y., and Lopez-Paz, D. (2019). Interpolation consistency training for semi-supervised learning. *Neural Networks Elsevier*.

[213] Viola, P., Jones, M. J., and Snow, D. (2005). Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161.

[214] Wan, J. and Chan, A. (2020). Modeling noisy annotations for crowd counting. *Advances in Neural Information Processing Systems*, 33.

[215] Wan, J., Liu, Z., and Chan, A. B. (2021). A generalized loss function for crowd counting and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1974–1983.

[216] Wang, G., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., Yu, G., Zhou, E., and Sun, J. (2020a). High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6449–6458.

[217] Wang, G., Yuan, Y., Chen, X., Li, J., and Zhou, X. (2018a). Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282.

[218] Wang, M., Cai, H., Dai, Y., and Gong, M. (2023). Dynamic mixture of counter network for location-agnostic crowd counting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 167–177.

[219] Wang, Q. and Breckon, T. P. (2022). Crowd counting via segmentation guided attention networks and curriculum loss. *IEEE Transactions on Intelligent Transportation Systems*.

[220] Wang, Q., Li, W., and Gool, L. V. (2019). Semi-supervised learning by augmented distribution alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1466–1475.

[221] Wang, S., Guan, Y., and Shao, L. (2020b). Multi-granularity canonical appearance pooling for remote sensing scene classification. *IEEE Transactions on Image Processing*, 29:5396–5407.

[222] Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. (2021a). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578.

[223] Wang, X., Girshick, R., Gupta, A., and He, K. (2018b). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803.

[224] Wang, Z., Li, X., Duan, H., Su, Y., Zhang, X., and Guan, X. (2021b). Medical image fusion based on convolutional neural networks and non-subsampled contourlet transform. *Expert Systems with Applications*, 171:114574.

[225] Wedel, M. and Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations*. Springer Science & Business Media.

[226] Wei, L., Zhang, S., Gao, W., and Tian, Q. (2018). Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88.

[227] Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

[228] Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., and Zhang, L. (2021). Cvt: Introducing convolutions to vision transformers. pages 22–31.

[229] Xia, B. N., Gong, Y., Zhang, Y., and Poellabauer, C. (2019a). Second-order non-local attention networks for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3760–3769.

[230] Xia, B. N., Gong, Y., Zhang, Y., and Poellabauer, C. (2019b). Second-order non-local attention networks for person re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*.

[231] Xiao, T., Liu, Y., Zhou, B., Jiang, Y., and Sun, J. (2018). Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434.

[232] Xie, Q., Dai, Z., Hovy, E., Luong, M.-T., and Le, Q. V. (2019). Unsupervised data augmentation for consistency training. *The Conference on Neural Information Processing Systems (NeurIPS)*.

[233] Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. (2020a). Unsupervised data augmentation for consistency training. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6256–6268. Curran Associates, Inc.

[234] Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. (2020b). Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698.

[235] Xiong, H., Lu, H., Liu, C., Liu, L., Cao, Z., and Shen, C. (2019). From open set to closed set: Counting objects by spatial divide-and-conquer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8362–8371.

[236] Xu, C., Qiu, K., Fu, J., Bai, S., Xu, Y., and Bai, X. (2019). Learn to scale: Generating multipolar normalized density maps for crowd counting. In *The IEEE International Conference on Computer Vision (ICCV)*.

[237] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

[238] Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678.

[239] Yang, J., Nguyen, M. N., San, P. P., Li, X., and Krishnaswamy, S. (2015). Deep convolutional neural networks on multichannel time series for human activity recognition. In *Ijcai*, volume 15, pages 3995–4001. Buenos Aires, Argentina.

[240] Yang, S., Guo, W., and Ren, Y. (2022). Crowdformer: An overlap patching vision transformer for top-down crowd counting. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, Vienna, Austria*, pages 23–29.

[241] Yang, Y., Li, G., Wu, Z., Su, L., Huang, Q., and Sebe, N. (2020). Reverse perspective network for perspective-aware object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4374–4383.

[242] Yao, L., Nie, F., Sheng, Q. Z., Gu, T., Li, X., and Wang, S. (2016). Learning from less for better: semi-supervised activity recognition via shared structure discovery. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 13–24.

[243] Yuan, K., Guo, S., Liu, Z., Zhou, A., Yu, F., and Wu, W. (2021a). Incorporating convolution designs into visual transformers. pages 579–588.

[244] Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.-H., Tay, F. E., Feng, J., and Yan, S. (2021b). Tokens-to-token vit: Training vision transformers from scratch on imagenet. pages 558–567.

[245] Yuan, Y., Chen, X., and Wang, J. (2020). Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer.

[246] Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032.

[247] Zaki, M. H. and Sayed, T. (2017). Automated analysis of pedestrian group behavior in urban settings. *IEEE Transactions on Intelligent Transportation Systems*, 19(6):1880–1889.

[248] Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., and Yang, M.-H. (2022). Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739.

[249] Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

[250] Zeng, M., Yu, T., Wang, X., Nguyen, L. T., Mengshoel, O. J., and Lane, I. (2017). Semi-supervised convolutional neural networks for human activity recognition. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 522–529. IEEE.

[251] Zeng, X., Guo, Q., Duan, H., and Wu, Y. (2021). Multi-level features extraction network with gating mechanism for crowd counting. *IET Image Processing*.

[252] Zeng, X., Wu, Y., Hu, S., Wang, R., and Ye, Y. (2020). Dspnet: deep scale purifier network for dense crowd counting. *Expert Systems with Applications*, 141:112977.

[253] Zhai, B., Perez-Pozuelo, I., Clifton, E. A., Palotti, J., and Guan, Y. (2020). Making sense of sleep: Multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(2):1–33.

[254] Zhang, C., Li, H., Wang, X., and Yang, X. (2015). Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 833–841.

[255] Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

[256] Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2019). Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR.

[257] Zhang, H., Long, Y., Guan, Y., and Shao, L. (2018). Triple verification network for generalized zero-shot learning. *IEEE Transactions on Image Processing*, 28(1):506–517.

[258] Zhang, Y., Zhou, D., Chen, S., Gao, S., and Ma, Y. (2016a). Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597.

[259] Zhang, Y., Zhou, D., Chen, S., Gao, S., and Ma, Y. (2016b). Single-image crowd counting via multi-column convolutional neural network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[260] Zhang, Z., Lan, C., Zeng, W., Jin, X., and Chen, Z. (2020). Relation-aware global attention for person re-identification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 3186–3195.

[261] Zhao, N., Chua, T.-S., and Lee, G. H. (2020a). Sess: Self-ensembling semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11079–11087.

[262] Zhao, T., Nevatia, R., and Wu, B. (2008). Segmentation and tracking of multiple humans in crowded environments. *IEEE transactions on pattern analysis and machine intelligence*, 30(7):1198–1211.

[263] Zhao, Z., Shi, M., Zhao, X., and Li, L. (2020b). Active crowd counting with limited supervision. *European conference on computer vision*.

[264] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*.

[265] Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., and Torralba, A. (2019a). Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321.

[266] Zhou, K., Yang, Y., Cavallaro, A., and Xiang, T. (2019b). Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3702–3712.

[267] Zhu, F., Shao, L., Xie, J., and Fang, Y. (2016). From handcrafted to learned representations for human action recognition: a survey. *Image and Vision Computing*, 55:42–52.

[268] Zhu, K., Guo, H., Liu, Z., Tang, M., and Wang, J. (2020). Identity-guided human semantic parsing for person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 346–363. Springer.

[269] Zhu, L., Zhao, Z., Lu, C., Lin, Y., Peng, Y., and Yao, T. (2019). Dual path multi-scale fusion networks with attention for crowd counting. *arXiv preprint arXiv:1902.01115*.