Imperial College
London

PARAMETRIC ACTIVE LEARNING TECHNIQUES FOR 3D HAND POSE ESTIMATION

RAZVAN CARAMALAU

Thesis submitted for the degree of Doctor of Philosophy

Supervised by DR TAE-KYUN KIM

Intelligent Systems and Networks group
Department of Electrical and Electronic Engineering
Imperial College London

November 2022

COPYRIGHT DECLARATION

ORIGINALITY DECLARATION

I hereby declare that this thesis and the work herein detailed, was composed and originated by myself, except where appropriately referenced and credited.

London, November 2022

Razvan Caramalau

ABSTRACT

Active learning (AL) has recently gained popularity for deep learning (DL) models due to efficient and informative sampling, especially when the models require large-scale datasets. The DL models designed for 3D-HPE demand accurate and diverse large-scale datasets that are time-consuming, costly and require experts. This thesis aims to explore AL primarily for the 3D hand pose estimation (3D-HPE) task for the first time.

The thesis delves directly into an AL methodology customised for 3D-HPE learners to address this. Because predominantly the learners are regression-based algorithms, a Bayesian approximation of a DL architecture is presented to model uncertainties. This approximation generates data and model-dependent uncertainties that are further combined with the data representativeness AL function, CoreSet, for sampling. Despite being the first work, it creates informative samples and minimal joint errors with less training data on three well-known depth datasets.

The second AL algorithm continues to improve the selection following a new trend of parametric samplers. Precisely, this is proceeded task-agnostic with a Graph Convolutional Network (GCN) to offer higher order of representations between labelled and unlabelled data. The newly selected unlabelled images are ranked based on uncertainty or GCN feature distribution.

Another novel sampler extends this idea and tackles encountered AL issues, like cold-start and distribution shift, by training in a self-supervised way with contrastive learning. It shows leveraging the visual concepts from labelled and unlabelled images while attaining state-of-the-art results.

The last part of the thesis brings prior AL insights and achievements in a unified parametric-based sampler proposal for the multi-modal 3D-HPE task. This sampler trains multi-variational auto-encoders to align the modalities and provide better selection representation. Several query functions are studied to open a new direction in deep AL sampling.

ACKNOWLEDGEMENTS

Primarily, I would like to send deep appreciation to my supervisor, Dr T-K Kim, for offering me this opportunity and for the ongoing valuable support. The research discussions conducted together have pushed me beyond my initial expertise and limitations.

On a second note, I would also like to exclusively thank my friend and post-doc within the ICVL group, Binod, who carried and shaped my research closely throughout my PhD. My warm gratitude toward him will continue to be manifested on my future career.

I would also like to acknowledge my colleagues in the ICVL lab. The long chats with my year colleagues have made the PhD experience more enlightening. Particularly, I want to mention Anil, who also played a meaningful part in my first year.

My warm appreciations go to my 1008D friends Dafni, Yang, Sara, Pamela, Ju-Il and Jonny. We shared multiple enjoyable and exciting lunches, coffee breaks, events and even birthdays. Their emotional support and shared struggles kept me on this 4-year adventure.

I also want to thank two important groups: Isha Foundation and the GradPad Griffon community. Isha Foundation has played a transforming part in my physic, mentality and emotions through their yoga and meditation practices. On the other hand, GradPad Griffon has been my home these years with great support from the staff and the Resident Assistants. I was lucky to meet not only so many wonderful people but also my girlfriend, Qiling. I also thank her for sharing the beautiful moments so far.

I also acknowledge the sponsoring of Huawei Research for this PhD.

Last but not least, a tremendous appreciation goes to my family, for which none of those things could have happened. I don't think I will be able to repay their support and gratitude in this lifetime, but I will try my best.

CONTENTS

1 INTRODUCTION 21
- 1.1 Motivation . . . 25
- 1.2 Problems to address and investigate . . . 28
- 1.3 The roadmap of the contributions . . . 29

2 BACKGROUND 33
- 2.1 Active Learning . . . 33
  - 2.1.1 Definition . . . 33
  - 2.1.2 Scenarios and categorisation . . . 34
  - 2.1.3 Initial selection methods . . . 35
  - 2.1.4 Deep Active Learning . . . 36
- 2.2 Theoretical foundations of the proposed Active Learning methods 37
  - 2.2.1 Uncertainties for vision applications . . . 37
  - 2.2.2 Graph Convolutional Networks . . . 38
  - 2.2.3 Self-supervised learning . . . 40
  - 2.2.4 Variational auto-encoder and Product of Experts . . . 41
- 2.3 3D Hand Pose Estimation . . . 43
  - 2.3.1 Definition . . . 43
  - 2.3.2 Categories . . . 44
  - 2.3.3 Deep Learning solutions . . . 45
- 2.4 Datasets for Hand Pose Estimation . . . 46
  - 2.4.1 Dataset Engineering . . . 46
  - 2.4.2 Depth image datasets . . . 46
  - 2.4.3 Multi-modal datasets . . . 47

3 ACTIVE LEARNING FOR BAYESIAN 3D HAND POSE ESTIMATION 49
- 3.1 Course of Objectives with Justification . . . 49
- 3.2 Relevant Literature . . . 50
  - 3.2.1 3D Hand Pose Estimation . . . 50
  - 3.2.2 Bayesian Deep Learning . . . 51
  - 3.2.3 Active Learning Methods . . . 52
- 3.3 Methodology . . . 53
  - 3.3.1 Bayesian 3D Hand Pose Estimator . . . 53
  - 3.3.2 Active Learning Framework . . . 56
- 3.4 Experiments . . . 59

3.4.1 3D Hand Pose Estimation Datasets . . . 59
3.4.2 Bayesian DeepPrior Architecture . . . 60
3.4.3 Active Learning Evaluation . . . 61
3.5 Conclusions and Limitations to address . . . 67

4 SEQUENTIAL GCN FOR ACTIVE LEARNING 69
4.1 Course of Objectives with Justification . . . 69
4.2 Relevant Literature . . . 70
4.2.1 Model-based methods . . . 70
4.2.2 Graph Convolution Network (GCN) in Active Learning 70
4.2.3 Uncertainty-based Active Learning methods . . . 71
4.2.4 Geometric-based Active Learning methods . . . 71
4.3 Method . . . 72
4.3.1 Learner . . . 72
4.3.2 Sampler . . . 73
4.4 Experiments . . . 77
4.4.1 Implementation details . . . 77
4.4.2 Quantitative comparisons . . . 79
4.4.3 Qualitative comparisons . . . 82
4.4.4 Ablation studies . . . 85
4.4.5 Sub-sampling of Synthetic Data . . . 87
4.4.6 3D Hand Pose Estimation . . . 88
4.5 Limitations and Conclusions . . . 90

5 LEVERAGING UNLABELLED DATA FOR ACTIVE LEARNING SELECTION 93
5.1 Course of Objectives with Justification . . . 93
5.2 Relevant Literature . . . 94
5.2.1 Self-supervised Learning . . . 94
5.2.2 Active Learning (AL) with self-supervision . . . 95
5.3 Methodology . . . 96
5.3.1 Contrastive Semi-supervised Learning framework . . . 96
5.3.2 Unlabelled samples selection . . . 99
5.4 Experiments . . . 100
5.4.1 Experiment settings . . . 100
5.4.2 Quantitative experiments . . . 101
5.4.3 Distribution shift discussion . . . 106
5.4.4 Self-supervised learning (SSL) modules variation and Ablation Study . . . 106
5.4.5 Selection function analysis . . . 107

5.4.6 SSL results and multi-stage AL . . . 109
5.5 Conclusions, Limitations and Future Work . . . 110

6 MULTI-MODAL VARIATIONAL AUTO-ENCODER FOR ACTIVE LEARNING 111
6.1 Course of objectives with justification . . . 111
6.2 Relevant Literature . . . 113
6.2.1 Multi-modal 3D Hand Pose Estimation (3D-HPE) . . . 113
6.2.2 Multi-modal AL . . . 113
6.3 Methodology . . . 114
6.3.1 Multi-modal 3D-HPE learner . . . 114
6.3.2 MVAAL . . . 115
6.4 Experiments . . . 119
6.4.1 Quantitative experiments . . . 119
6.4.2 MVAAL Qualitative Analysis . . . 122
6.5 Future work and Conclusions . . . 123

7 CONCLUSIONS AND FUTURE WORK 125
7.1 Contributions and Achievements of the Thesis 125
7.2 Limitations and Challenges 127
7.3 Future work 128
7.4 Statement from the author 129

BIBLIOGRAPHY 131

A ACTIVE LEARNING QUALITATIVE ANALYSIS FOR 3D HAND POSE ESTIMATION 147

B EXTENDED EXPERIMENTS FOR THE GCN-BASED SAMPLER 149
B.1 CIFAR-10 imbalanced dataset 149
B.2 Ablation study - GCN parameter search 150
B.3 VGG-11 learner for CIFAR-10 image classification for 3 selection stages 150
B.4 Extended qualitative analysis on the AL method 151

C PUBLICATIONS OF THE AUTHOR 153

LIST OF FIGURES

Figure 1.1 Roadmap of the thesis chapters parallelised between AL and 3D-HPE. 29

Figure 2.1 A typical Deep Active Learning (DAL) cycle. It starts with the initial training of labelled examples. Then, feature maps may be required by the AL sampler from the model. Once selected, the new batch is sent to the oracle for labelling. Finally, the training set is updated for another cycle. 36

Figure 3.1 Schematic diagram showing the end-to-end pipeline of the AL method. The depth image of the hand is pre-processed and fed to the Bayesian DeepPrior Network. This network is trained to minimise the objective function given in Equation 3.2. The aleatoric and epistemic uncertainties of a sample that we obtain from M number of Monte Carlo Dropouts are passed on to the selection criteria for querying unlabelled examples. The selection method is highlighted in the Algorithm 1. The switches in the Figure are present to indicate the activation of Dropout in different Bayesian approximations. 51

Figure 3.2 Conceptual comparison between the sampling techniques: Corset (Sener and Savarese, 2018) vs CKE. The blue dots represent the annotated samples in the hand skeleton space, while the red ones are to be selected. Each sample's uniform distribution defines the circle radii in the CKE process stages. 58

Figure 3.3 Empirical comparison I: Quantitative analysis of the proposed CKE method with Bayesian DeepPrior against the other methods applied on the standard version. Evaluated datasets: ICVL (Top), BigHand2.2M (Middle) and NYU (Bottom). 62

Figure 3.4 Empirical comparison II: AL performance comparison of the proposed CKE method against other Monte Carlo (MC) Dropout adapted samplers on ICVL(Top), BigHand2.2M(Middle) and NYU(Bottom) data set. . . 63
Figure 3.5 Qualitative comparisons on the Active Learning methods at different selection stages (Left). The 3D joint aleatoric and epistemic uncertainty variance (Right). . 66
Figure 4.1 The diagram outlines the proposed pipeline. **Phase I**: Train the learner to minimise the objective of downstream task from the available annotations, **Phase II**: Construct a graph with the representations of images extracted from the learner as vertices and their similarities as edges, **Phase III**: Apply GCN layers to minimise binary cross-entropy between labelled and unlabelled, **Phase IV**: Apply uncertainty sampling to select unlabelled examples to query their labels, **Phase V**: Annotate the selection, populate the number of labelled examples and repeat the cycle. 72
Figure 4.2 This Figure simulates the sampling behaviour of the **UncertainGCN** sampler at its first three selection stages. A toy experiment is run just by taking 100 examples from the ICVL (Tang et al., 2014) hand pose dataset. Each node is initialised by the features extracted from the learner, and the edges capture their relation. Each concentric circle represents a cluster of strongly connected nodes. Here, a group of images having similar viewpoints are in a concentric circle. Considering two annotated examples as seed examples in the centremost circle, in the first selection stage, the algorithm selects samples from another concentric circle which is out-of-distribution than selecting the remaining examples from the innermost circle. Similarly, in the second stage, the sampler chooses images residing in another outer concentric circle which are sufficiently different from those selected before. 74
Figure 4.3 Quantitative comparison on CIFAR-10(Top) and CIFAR-100(Bottom) 80
Figure 4.4 Quantitative comparison on FashionMNIST(Top) and SVHN(Bottom) 81

Figure 4.5 Exploration comparison on CIFAR-10 between CoreSet and UncertainGCN at the first (Top) and fourth (Bottom) AL selection stage . . . 83
Figure 4.6 Exploration comparison on CIFAR-10 between CoreSet and CoreGCN at the first (Top) and fourth (Bottom) AL selection stage . . . 84
Figure 4.7 UncertainGCN tuning parameters . . . 86
Figure 4.8 Learner comparison [Resnet-18 vs VGG-11] - 4,000 CIFAR-10 labelled images [mean averaged accuracy on 5 trials] . . . 87
Figure 4.9 Performance comparison on sub-sampling synthetic data to augment real data for expression classification 88
Figure 4.10 Quantitative comparison on 3D Hand Pose Estimation - BigHand2.2M(Top), NYU(Middle), ICVL(Bottom) (lower is better) . . . 89
Figure 5.1 SSL-AL training framework under the proposed MoBYv2 configuration. The query feature encoder plays two roles: to map the features to the task discriminator for classification; to capture contrastive visual representation with the asymmetry of the query and key modules. For unlabelled data, the blue lines show the back-propagation of contrastive loss and its exponential moving average (dashed). The green lines also include the cross-entropy loss during training when the annotation is available. . . . 95
Figure 5.2 Evaluations on CIFAR-10 (Top), CIFAR-100 (Bottom) . 103
Figure 5.3 Evaluations on SVHN (Top), FashionMNIST (Bottom) . 104
Figure 5.4 CIFAR-10 imbalanced dataset experiment(Top); Mitigating the distribution shift with MoBYv2(Bottom) . . . 105
Figure 5.5 Quantitative evaluation with different selection functions for CIFAR-10 (Top), CIFAR-100 (Bottom) . . . 108
Figure 5.6 Qualitative AL selection analysis on MoBYv2. t-SNE representations at the first selection stage for five classes of SVHN . . . 109

Figure 6.1 The proposed AL pipeline: the multi-modal 3D-HPE learner (Top); MVAAL sampler (Bottom) with selection functions (Right); Two modalities are represented in the figure, but more can be added. MVAAL constitutes of the pre-trained multi Variational Auto-Encoders (VAE) with the Product of Experts (PoE) alignment; the pose discriminator and the selection function (as sampler variants). MVAAL fine-tunes the pose discriminator with the new labelled subset that was assigned to the learner after each AL cycle. 116
Figure 6.2 Depth-maps reconstructed with the multi-modal VAE at different training points. 122
Figure 6.3 Multi-modal VAE reconstructed modalities from the PoE 123
Figure A.1 Qualitative AL selection. Comparison between entropy uncertainty and CoreSet according to viewpoint and articulation distributions on BigHand2.2M 147
Figure A.2 Qualitative AL selection. CKE selected samples according to viewpoint and articulation distributions on BigHand2.2M 148
Figure B.1 Quantitative results - CIFAR-10 imbalanced dataset . . 149
Figure B.2 Ablation studies - CIFAR-10 GCN Hyper-parameters tuning 150
Figure B.3 CIFAR-10 Learner VGG-11 - 3 selection stages 151
Figure B.4 Extended qualitative analysis on labelled/unlabelled images at the last selection stage for CIFAR-10, ICVL and RaFD 151

LIST OF TABLES

Table 3.1 Ablation evaluation of Bayesian DeepPrior 60
Table 3.2 Bayesian DeepPrior vs DeepPrior - averaged mean squared error (MSE) [mm] 61
Table 3.3 Averaged epistemic and aleatoric variances at $\times 10^{-2}$ on ICVL, NYU and BigHand2.2M 61
Table 5.1 Comparison with the SSL-AL method CSAL on CIFAR-100 with a WideResNet-28 learner 102

Table 5.2 Variation of SSL pipeline. Average testing performance (5 trials) on CIFAR-10 for 3 AL cycles with ResNet-18 encoder . . . 107
Table 5.3 Ablation study of MoBYv2. Average testing performance (5 trials) on CIFAR-10 for 3 AL cycles with ResNet-18 encoder . . . 107
Table 5.4 Multi-stage SSL-AL vs Jointly end-task AL.Testing performance on CIFAR-10 with ResNet-18 encoder . . 109
Table 5.5 Semi-supervised learning comparison. Testing performance on CIFAR-10 with ResNet-18 encoder 109
Table 6.1 Quantitative AL results on STB dataset. Averaged joints MSE over 5 trials. % of annotated data 120
Table 6.2 Quantitative AL results on RHD dataset. Averaged joints MSE over 5 trials. % of annotated data 121
Table 6.3 Quantitative AL results on FreiHand dataset. Averaged joints MSE over 5 trials. % of annotated data . . . 121
Table 6.4 Quantitative AL results on RHD dataset. RGB Single Modality. Averaged joints MSE over 5 trials. % of annotated data 122

LIST OF ALGORITHMS

Figure 1 CKE 58
Figure 2 UncertainGCN active learning algorithm 76

ACRONYMS

3D-HPE 3D Hand Pose Estimation

Adam Adaptive Moment Estimation

AI Artificial Intelligence

AR Augmented Reality

AL Active Learning

BNN Bayesian Neural Network

CPMs Convolutional Pose Machines

CNNs Convolutional Neural Networks

CV Computer Vision

DOF Degree(s) of Freedom

DL Deep Learning

DAL Deep Active Learning

EBF Energy-based function

ELBO evidence lower bound

GANs Generative Adversarial Networks

GCN Graph Convolution Network

GPUs Graphical Processing Units

HCI human-computer interaction

KL Kullback-Leibler

MC Monte Carlo

ML Machine Learning

MCP metacarpophalangeal

PIP proximalcarpophalangeal

DIP distalcarpophalangeal

MLP Multi-Layer Perceptron

MSE mean squared error

NLP Natural Language Processing

PCA Principal Component Analysis

PSO Particle Swarm Optimisation

PoE Product of Experts

QBC query-by-committee

RGB-D Red Green Blue and Depth

RL Reinforcement Learning

SLAM Simultaneous Localisation and Mapping

SGD Stochastic Gradient Descent

SSL Self-supervised learning

t-SNE t-distributed stochastic neighbor embedding

VAE Variational Auto-Encoders

VAAL Variational Adversarial Active Learning

VR Virtual Reality

1

INTRODUCTION

The four years of this thesis are placed in the times of ongoing development of the *Deep Learning (DL)* field. This branch of *Machine Learning (ML)* has enabled artificial intelligent (*Artificial Intelligence (AI)*) systems to surpass humans in various types of tasks. While the first concepts of neural networks as ensembles with hidden layers date from the 1980s Rumelhart et al. (1986), it was not until 2006 that Hinton et al. (2006) proposed layer-wise training of similar models for deep belief networks. This greedy-learning algorithm has shown to be theoretically promising, especially with the uprising of large public datasets like ImageNet (2009) Deng et al. (2009).

However, the starting practical milestone of DL is widely recognised in Krizhevsky et al. (2012) for the image classification task. The work of Krizhevsky has standardised the Convolutional Neural Networks (CNNs) as a robust differentiable shared structure (with pre-defined concepts from the neocognitron Fukushima (2004)) for state-of-the-art visual learnt representations. Moreover, the proposed open-source framework has enabled current researchers further to extend the CNNs to various computer vision applications. Also, this may not have been possible without the scalability and deployment of Graphical Processing Units (GPUs) training and testing.

These recognisable pillars of DL consolidate the *area of study* for this thesis with a focus on computer vision applications. Although the technical and deployment advancements have contributed tremendously, the extension to other ML tasks has been enabled with the apparition of dedicated datasets.

To form a *dataset*, there are several steps to consider. A team has first to address the purpose or the scope of gathering the data. Then, under this *objective* comes a planning stage where the acquisition and annotation parts are configured together with the remaining parameters. Once established, the *data acquisition* process can start within a specific observing environment with a corresponding capturing system. With each acquired data, the *annotation system* can provide meta-data or targeted information in alignment with the objective. The annotation system may consist of a human oracle (in a manual procedure), an automated label generator or a semi-automated combination of those. Finally, the dataset culminates with a testing and validation stage.

This step assures that the gathered data and the registered corresponding information fit the initial planning and purpose.

Prior to the work conducted in the thesis, there has been a substantial disparity between the papers that focus on DL algorithms and the ones that extend the research in datasets. It is a common trend that a trail of new algorithms would appear while outputting incremental performance with the release of a new dataset. In many cases, this behaviour would neglect the datasets' quality, which would further become reference benchmarks for many applications. Taking the image classification task, the ImageNet dataset has been widely recognised for great complexity and class diversity. However, several issues have been reported with class mislabelling. This aspect has been widely observed in other datasets as well Northcutt et al. (2021).

From another perspective, a variety of datasets tend to constitute, in time, redundant data regarding the advanced learning capabilities of newer DL models. This behaviour has been shown gradually in datasets like MNIST LeCun and Cortes (2010), CIFAR-10/100 Krizhevsky (2012), SVHN Goodfellow et al. (2013*a*) and many more. The advanced DL models achieved a great level of generalisation over these datasets, encouraging researchers to extend their pipelines to real-world environments. As a result, the gaps between standard benchmarks, real-world scenarios, and specific DL architecture have drastically increased. E.g., assuming that specific models can generalise well on public data does not simply guarantee the same convergence over newly acquired datasets. In assisting to bridge this gap, researchers have started since 2016 to revisit classic ML approaches like *Active Learning (AL)*.

AL, in standard practice, can be defined as a recurrent dataset formation containing the learning algorithm (also known as *learner*) in the loop. Keeping the learner is to find the most informative and meaningful samples from where it could further improve. This mainly defines the objective of AL. Therefore, the key ingredient of AL stands on developing the selection function strategy (broadly known as *sampler*). If the AL tracks the optimal data for the learner's performance, then it is expected to reduce the number of required samples considerably. As a matter of fact, AL can also be applied for pruning the available data, even before the annotation process. Reducing the labelling effort presents a clear advantage compared to unguided sampling, and this may be supportive in the planning stage of the dataset formation. In the majority of cases, the annotation stage is known to be laborious, time-consuming and expensive. To limit the noise and error deviations from ideal ground truth, experts must provide high-quality labels. Taking these condi-

tions, several applications have included AL as an essential part of creating and updating their datasets.

Apart from computer vision, domains that have benefited from the AL strategies vary from medical imaging and Augmented Reality (AR) for education to Natural Language Processing (NLP) and the automotive industry.

Companies like NVIDIA Haussmann et al. (2020), Tesla, and Waymo have been adding AL in their pipelines to find and interact with the failures of the perception systems. Data efficiency through active learning has mitigated the effect of the long-tail distribution that occurs in automotive recording systems. From another perspective, medical applications rely on credence to arrive at clinical decisions. Here, AL plays the role of a decision refiner with the human(oracle)-in-the-loop advantage Budd et al. (2021).

In NLP, tasks like identification, completion and recognition adopted a more effortless extension to large-scale datasets by integrating AL. Moreover, specific measurements and uncertainties introduced with the selection strategy allowed further insights from the "black-box" caveat of DL models Luo et al. (2021). Education and practical training through AR has recently revolutionised classical teaching. These immersive technologies offer students an interactive and valuable experience while reducing material and physical costs. The range of tasks, however, is constrained by the simulated environment. Together with this simulator, DL-based infrastructures are deployed with AL to assist the user in 3D scene reconstruction, Simultaneous Localisation and Mapping (SLAM) and pose estimations Jesionkowska et al. (2020).

Following the application domain, the thesis centres the focus of AL strategies on vision tasks. Linked to the AR for educative purposes, but not limited to it, 3D Hand Pose Estimation (3D-HPE) is the primary task to be investigated. Therefore, the *research topic* of the thesis stands at the congregation of AL and 3D-HPE. Before delving into the motivation of this topic, it is supportive of addressing the objective and area of the 3D-HPE problem.

Estimating the pose is fundamental in human-machine interaction nowadays. Therefore, a robust model of the human is needed to map the locations of the body, head and hand. In this model, the kinematics and postures have to be realistic and contain all the Degree(s) of Freedom (DOF) a human can express. For hand pose estimation, the taxonomy established in the survey from Erol et al. (2007) follows the 27-bone structure of the hand. This hand structure consists in 21 joints of interest, where the root is considered at the fixed base of the palm. Six of the joints, from the palm, have two DOF while

the rest of the joints to the fingertips only one each. Thus, by determining the coordinates of the joint locations between the moving bones, the location of the hand can be reconstructed.

The 3D-HPE assumes an annotated hand model where all the distributed key points, commonly taken from joints, have the 3D plane coordinates. Training a learning algorithm to identify these locations from an environment defines the paradigm of 3D-HPE.

In terms of environment, the inputs usually are recorded RGB or depth image frames from a camera-sensor system. Although RGB or grey cameras are relatively cheap and easy to deploy for acquisition, the annotation and the pose estimation can present several challenges. For setting up the accurate 3D hand ground truth, external sensors with control systems or hand modelling are required Hsiao et al. (2016); Kato and Takemura (2016). Despite this, further refinement of the oracle from the RGB/grey data would still end in noisy annotations. While overcoming external perturbations like lighting, noise or brightness variations, depth cameras have become a principal component in strengthening the quality of annotations and input data. Therefore, nowadays, datasets often get formed from multi Red Green Blue and Depth (RGB-D) camera systems Moon et al. (2020); Nie et al. (2019).

3D-HPE is at the base of diverse human-computer interaction (HCI) applications. Along with the education field mentioned before, the entertainment industry has gained another level of interaction, beginning with Microsoft's Kinect depth sensor for games. Recently, the technology evolved to AR/Virtual Reality (VR) goggles and consoles that surround the user in a modelled reality, commonly referred to as meta-verse. This is an ongoing extension domain of real-world applications. Games, video-conferencing, AR shopping and real estate viewing are just a few of them to date. When referring to the robotics industry, domestic or industrial, the 3D-HPE plays an even bigger role where learning algorithms must produce precise motions, actions and gestures. Domestic robots would require to exert several hand-object actions for activities like ironing or cleaning. On the industrial side, robots are heavily used in manufacturing, and the control systems should produce robust manipulations. This is also the case in the medical area, where surgical robots are successfully supervised on diverse operations.

Recognising the relevance of both AL and 3D-HPE research areas, the thesis tackles for the first time the interactions and challenges that these communities have recently faced. Throughout the thesis, AL and 3D-HPE will be treated in a combined manner with parallel investigations in the AL sampling strategies. For simplicity, some of the AL proposed framework will

be deployed in a task-agnostic manner. This will allow an undemanding experimental flow on tasks such as image classification.

After the introduction of the thesis's topic and research area, the next sections will further elaborate on the derived *motivations* together with the *roadmap* of the upcoming chapters.

1.1 MOTIVATION

Previously and over the timeline of the thesis, DL has become more and more dominant as ML solutions to a wide range of applications. In definition, DL forms representation learning functions with the deep innovative architectures of neural networks. The most successful stories have started with the deployment of supervised training on CNNs Krizhevsky et al. (2012); Simonyan and Zisserman (2015). This, together with the ongoing development, is due to three main factors: large-scale datasets; efficient GPUs computation for training and testing; open-source code and optimal frameworks for implementation like Caffe, Tensorflow, and Pytorch.

DL models come with great scalability showing that with more data, better concepts are learned, and as an outcome, the performance gradually improves. In particular, the CNNs produce better visual concepts than standard Computer Vision (CV) feature descriptors. This is due to the numerous shared parameters of the spatial kernels over the entire dataset.

3D-HPE with the DL breakthrough. State-of-the-art models for 3D-HPE prior to CNNs were constituent of random forests pipelines (Tang et al. (2014); Tang et al. (2013)). Although both types of architectures can easily adapt to large datasets, the CNNs-based ones have been dominating in performance and deployment since 2015 (Baek et al. (2018*a*); Chang et al. (2018); Oberweger and Lepetit (2017); Oberweger et al. (2015); Ye et al. (2016)). As a learning objective for 3D-HPE, most of the methods follow an error minimisation procedure between predicted key points and the ground truth of the hand's position. Thus, the estimator can be interpreted as a highly non-linear regression function over these key points. In this context, the "black-box" behaviour of a trained CNN will limit any possible *confidence measurement* to unseen data. Moreover, after visual qualitative evaluation of predicted poses, back-tracking over *the model's decision chain* becomes even more challenging.

To address these two issues, specific to 3D-HPE, the thesis proposes a universal Bayesian approximation to DL models so that confidence measurements are generated to both seen and unseen data. This Bayesian adaptation

is inspired by Kendall and Gal (2017), where two types of uncertainties are captured from the model during training and testing. The uncertainties are measured for each input image through standard deviations of the predicted hand skeleton. To bring more insights into unseen data, one set of uncertainties will quantify hard or noisy samples, and another will evaluate the model's capability to improve further.

Lack of data representativeness for 3D-HPE. DL solutions are data-hungry also in the context of 3D-HPE. Therefore, from a dataset perspective, several RGB, depth and synthetic benchmarks have been widely explored. ICVL Tang et al. (2014), MSRC Sharp et al. (2015), and NYU Tompson et al. (2014) are some of the most popular datasets at the beginning of this thesis. It was recognised within the research community that depth-based datasets are still limited to global data distribution, given the variety of camera viewpoints and occluded possibilities. Consequently, synthetic hand-shapes can broadly fill this gap, however, they often present unrealistically articulations. This denotes that the DL model will learn inconsistent samples.

Researchers have addressed these shortcomings by extending the dataset sizes to over 2 million real samples. The first proposed was BigHand2.2M Yuan, Ye, Stenger, Jain and Kim (2017), followed later by InterHand2.6M Moon et al. (2020). Both datasets have been acquired samples from multiple viewpoints and with a high degree of articulation and hand shapes. Through the Hands In the Million challenge (Yuan, Ye, Garcia-Hernando and Kim, 2017*a*), which combined data of BigHand2.2M and First-Person Hand Action dataset (Garcia-Hernando et al., 2018), a study has been elaborated proving that most of the methods successfully generalised on the testing set Yuan et al. (2018). Acknowledging the advancements of the 3D-HPE methods, it was previously concluded that the systems still *do not generalise* enough for real-time applications Supancic et al. (2015).

On this matter, then the main concern arises in how many samples are required to increase robustness for 3D-HPE in the wild. An interesting experiment presented in Yuan, Ye, Stenger, Jain and Kim (2017) evaluates the performance of an estimator on randomly sampled subsets at different ratios of BigHand2.2M. The experiment shows little to no error when training even with a 16th of the entire dataset; this results in a lot of *redundancy within the samples*.

This thesis focuses on addressing these points through AL sampling. Many AL selection functions combine representativeness between the data with the model's uncertainty evaluation. Similarly, a novel sampler is explicitly

designed for 3D-HPE. It also includes the two types of uncertainties investigated and generated previously. The AL strategy aims to find which subsets of data can be more informative and simultaneously produce close accuracy to full dataset training.

AL for data efficiency. For DL, more data implicitly assumes better performance. For 3D-HPE, exploring diversity through viewpoints, articulations and hand shapes is a good strategy for better generalisation of the model. However, as it was noticed with BigHand2.2M, learning solutions may find some sequences of data repetitive or with marginal gain in the pose estimation. Classically, AL has been seen as an online learning environment where both the model and the oracle interact with live data in a loop. In this manner, first, the sampler selects new data, then the oracle labels it, and finally, the model is updated with the new data. This strategy is *unachievable in the context of DL* where models require a lengthy time to update many parameters.

In this thesis, AL is re-defined for DL to facilitate the large learning models. This area has been named Deep Active Learning (DAL), and it keeps the same focus on creating the most suitable samplers. However, the selection is made in an offline fashion where batches of images are stored from an unlabelled pool of data. This strategy can even be transferable to a data efficiency task for pruning the current labelled set in regard to a pre-trained model.

Limitations of DAL So far, the motivations have converged to a DAL framework for 3D-HPE by integrating two types of uncertainties and representativeness evaluation. This framework assumes an initial labelled training set from which the model can learn and further improve. The training set is often constituted from uniformly sampled data of the unlabelled pool. With no prior information regarding the pool of data and the model's ability to discriminate, the AL selection could be negatively influenced by the initial set. This shortcoming has been previously referred to as the *'cold-start' problem* Maltz and Ehrlich (1995); Schein et al. (2002). For DAL, the effect of 'cold-start' may even be detrimental when highly parameterised models are optimised.

Another limitation that DAL can pose is the *distribution shift*. This effect usually occurs when a new batch of data is added to the training set after an AL selection cycle. As a consequence, the re-trained model may shift its objective function from the current minima. The AL sampler's decision would implicitly be affected if this predominantly relies on the model's uncertainty.

To address these challenges, typically for AL, and refine the selection, the thesis diverges to a more simple vision task, specifically, image classification.

New AL frameworks will be designed in a task-agnostic way so that their acquisition functions can further expand to 3D-HPE. Both solutions will evaluate dependencies within the data while mitigating the effects of 'cold-start' and distribution shift.

1.2 PROBLEMS TO ADDRESS AND INVESTIGATE

The problems to address in this thesis fall in the DAL research area with a distinct interest in the 3D-HPE task. Therefore, these can be summarised as questions accordingly:

1. "Given a DL-based 3D hand pose estimator, how can the learner reach minimal error with AL when meaningful real sampled data are provided?"

2. "How can the acquisition function in DAL customise its selection in a more informative and task-aware manner?"

3. "How can self-supervision help AL in dealing with the 'cold-start' problem and distribution shift? Can the unlabelled examples be leveraged to improve the initial visual concepts of the learner?"

4. Hand pose estimators that approached multi-modalities during training have been the most successful. "How do these modalities influence the performance gain in the context of DAL?"

The first question is drawn from the lack of representativeness between the large-scale datasets of hand poses and the DL architectures. As a solution, AL is proposed for the first time on the 3D-HPE task.

The second question comes from the demand for better AL selection functions for DL models. A recent trend has opened a new direction of research where dedicated parameterised samplers can boost the diversity for DAL in the acquisition process. Thus, more investigations are derived from this approach.

Following the same trend in DAL, the third question revisits the standout caveats of this field. It aims to tackle the 'cold-start' problem and the distribution shift by increasing the visual representativeness with self-supervision over the entire available data.

The last question from the list revisits the first one but under a multi-modal learning setting. A new AL strategy is adapted to find the most informative sets of modalities from a teacher network. The teacher network

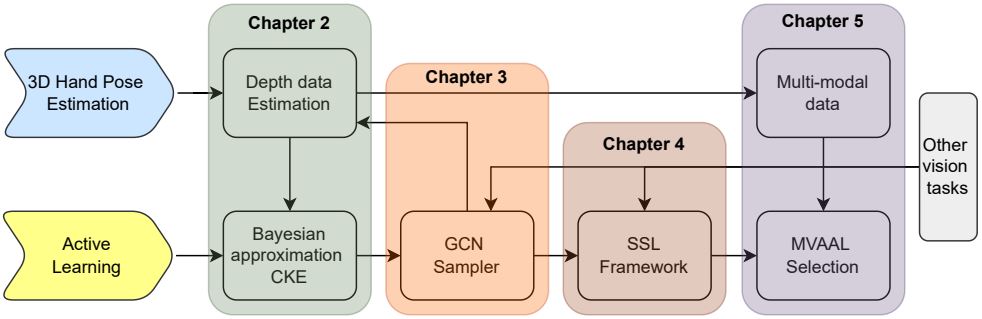

Figure 1.1: Roadmap of the thesis chapters parallelised between AL and 3D-HPE.

acts like a visual representation guide to the learner, so it can achieve greater generalisation with less number of samples.

1.3 THE ROADMAP OF THE CONTRIBUTIONS

In an attempt to answer the previous questions, the thesis unfolds three main contributions in the next chapters, followed by the current and future work. Figure 1.1 displays the two flows of research where AL methodologies intertwine with 3D-HPE and other vision tasks. The coloured blocks frame the contributions to the corresponding chapters as follows:

Chapter 3:

- It proposes for the first time to date the 3D-HPE task for depth image inputs under the AL framework.
- An approximation of Bayesian hand pose estimators is suggested to extract uncertainty measurements from the regression-type architectures.
- It presents a novel AL sampling method that incorporates both predicted skeletons and their corresponding uncertainty variances.
- It systematically evaluates the AL sampling cycles for both standard and the Bayesian approximated models on three well-known 3D-HPE benchmarks: BigHand2.2M (Yuan, Ye, Stenger, Jain and Kim, 2017), ICVL (Tang et al., 2014) and NYU (Tompson et al., 2014).
- The proposed solution consistently outperforms the counterpart competitive baselines quantitatively.

Chapter 4:

- It follows the new trend of parameter optimised AL samplers, and it shifts the objective of its selection in a task-agnostic way.
- It deploys a Graph Convolution Network (GCN) architecture as a sampler to model the long-range dependencies within the features space while trying to separate between labelled and unlabelled extracted features. It aims to propagate better informativeness for the learner and assist the selection criteria. This also addresses the limitation encountered in Variational Adversarial Active Learning (VAAL) (Sinha et al., 2019), a Variational Auto-Encoders (VAE)-based AL sampler, where its training is disconnected from the learner.
- Two selection criteria are studied along with the GCN sampler: one, data representative and another, uncertainty based. A further investigation into their suitability according to each vision task is presented.
- It evaluates the GCN sampler with its two selection criteria on four challenging image classification datasets, the same 3D-HPE benchmarks from Chapter 3 and a synthetic-generated face expression dataset.
- After these qualitative experiments and a qualitative study, the task-agnostic proposal shows consistent improvement over all the other AL techniques to date. This also enables developers to decide on the final selection criteria depending on their application.

Chapter 5:

- It puts the AL task in a self-supervised perspective. Therefore, the learner is wrapped in a key-query contrastive learning pipeline. The self-supervised architecture has similarities to one of the most recent methods, MoBY (Xie et al., 2021), however, it is trained jointly with the end task.
- It addressed the main issues of DAL, the 'cold-start' problem and the distribution shift.
- The Self-supervised learning (SSL) pipeline trains with both labelled and unlabelled so that enriched features can further optimise a data representative (geometric) AL selection. This proposed AL procedure is identified as MoBYv2AL.

- It tackles quantitative experiments mainly for image classification. Due to the nature of the SSL, it creates a high-performance gap on four datasets.

Chapter 6:

- It adopts the multi-modal input learners, in particular for 3D-HPE, for the AL problem. With the availability of more camera sensors and systems, multi-modal datasets have become a standard nowadays. Thus, a strong interest in making these systems more efficient needs to be addressed.

- For sampling, it upgrades the concepts of VAAL to multi-modality. Distinctively, the VAE sampler is trained in a unified generative framework with a product of experts as in Yang et al. (2019). This sampler, denoted as MVAAL, creates a shared latent space between the modalities. In exchange, this offers combined high-level representations for better decisions in the sample query process.

- It deploys several other baselines together with MVAAL for three multi-modal hand datasets: STB (Zhang et al., 2017), RHD (Zimmermann and Brox, 2017), and FreiHand (Zimmermann et al., 2019).

- It demonstrates through performance tests that the multi-modal learner benefits from different AL subsets of data. This motivates pursuing new future AL solutions or dealing with the learner more structurally.

In the end, **Chapter 7** brings together all the findings and concludes with future work. The chapter reiterates the relevance of the enclosed research and its impact on the AL community. To assist other researchers, this part also shares certain limitations of this thesis.

A list of appendices is gathered after the main chapters. The highlights of these are accordingly:

- **Appendix A** conducts an initial qualitative analysis of the two uncertainties described in Chapter 3.

- **Appendix B** adds supplementary information to the parameter settings of the GCN sampler from Chapter 4. It also displays the type of data used in the experiments.

- **Appendix C** specifies the published works that founded this thesis.

In the next **Chapter 2** of this thesis, a theoretical background is encompassed to solidify the understanding of the chapters listed above. This presentation brings the reader closer to the AL field and to the methodologies approached throughout the next sections.

2

BACKGROUND

The theoretical background is divided into four main parts:

- The first section 2.1 delves into the AL core principles. It presents the standard strategies and the selection methods, from classic ones to recent DAL approaches.
- Section 2.2 goes through the complementary theory of the methods designed in the next chapters.
- The third section defines 3D-HPE, the main application analysed for AL. It further details the progression of state-of-the-art DL architectures,
- The final part, 2.4, brings the dataset engineering structure for the current hand pose estimation datasets. This is essential for AL, especially in optimising and adding data.

2.1 ACTIVE LEARNING

2.1.1 *Definition*

The goal of AL consists in finding the most informative data for a given ML model. It is fundamental as well to achieve higher performance with less amount of data. Therefore, AL can be seen as a cyclical process, composed of four main elements: **input data space** (labelled/unlabelled dataset); **learner** (learning model / algorithm); **query function** (acquisition method / sampler); **oracle** (human-machine annotator). Each component can be defined as follows:

- The *input data space* is the observing environment from which instances are sampled. This can be through single or multiple instances. The data space is used by all the other elements of AL.
- A *machine learning model* defines the objective of the application and it is dependent on the supervised input data.
- The role of the *query function*, being at the core of AL, is to select hard samples from where the learner can effectively improve.

- The last part consists of the *oracle*, which contribution is to correctly annotate the previously selected samples. This will bring new samples to the labelled set, suitable for the learner.

The cyclical process starts from the input data space. First, instances are sampled and labelled if there is no prior data available. After this, the learner updates its function and parameters by supervised training on the labelled data. Then, new instances are selected from the input distribution according to the query function. Finally, the new queries are sent to the oracle for labelling so that the learner can start training again. In the majority of cases, the learner is actively used in the selection process.

2.1.2 *Scenarios and categorisation*

According to the survey from Settles (2009), AL has been identified in working under three scenarios. These can be appropriate depending on the type of learner and its dynamic with the input space.

Membership synthetic sampling is one of the first scenarios that initiated AL. It relies on randomly selecting data that was generated through a membership rule. Therefore, to avoid misrepresentation, similarly to Lang and Baum (1992), the quality of the samples should be analysed before selection. Although the randomness might yield in inefficient samples, this scenario would still be suitable where the annotation is difficult and noisy. Moreover, this scenario can be further improved if the membership rules are well established within the input space (Schumann and Rehbein, 2019).

In an attempt to address the randomness in the selection, the second scenario proposes a *selective sampling for streams* of inference (Cohn et al., 1994). The stream behaviour assumes that the learner is continuously exposed to instances and the AL decides whether to label or discard. On the other hand, the query function should be relatively simple and produce immediate decisions. Furthermore, the learner should be able to update its parameters with the newly selected samples to maintain the stream.

In the scenario where the learner cannot be trained online, the input instances are gathered in a pool of data. This is commonly referred as the *pool-based scenario* (Lewis and Gale, 1994). Like in the stream-based case, a query function is deployed, however, its process can use the entire pool over a longer time frame. Thus, batches of new instances are selected to be labelled with each cycle.

One way of categorising AL is by its selection principle of the query function. This can be broadly separated between functions that explore *uncer-*

tainty and others that use *data representativeness*. Another category fuses the measurements from both uncertainty and representativeness for its selective decision.

On the one hand, at the base of uncertainty sampling (Lewis and Gale, 1994) lies the concept of finding the least certain samples. This primarily tackles the most challenging samples for the learner. It will likely favour distribution outliers and areas close to decision boundaries.

On the other hand, the main principle of representative or geometric (density-weighted) (Settles and Craven, 2008) methods is sampling in clustered (dense) areas with the expectation of separateness. The instances from the underlying distribution are useful to increase the representativeness of the learner as a whole. The caveat in over-sampling from uncertain areas is the distribution shift. This may occur especially in the pool-based scenario where the update of the parameters happens over sets of data. Despite this, the density-weighted methods can implicitly overcome this issue.

More recently, depending on the structure of the AL sampling, this can be designed as *a selective function, an ensemble, a dedicated sampler or as an entire framework*. An AL function frequently uses a simple ranking criterion of certain measurements. From here, the structure can be extended to an ensemble of learners, also called query-by-committee (QBC) (Seung et al., 1992). These are trained altogether and derive their selection with a voting rule. With the popularity of DL training, researchers have shifted the AL selection as a loss minimisation problem for parameterised models. For example, these samplers would inherit uncertainty either by predicting the loss of the learner (Yoo and Kweon, 2019) or when trained in an adversarial manner (Sinha et al., 2019). The final structure can be viewed as a framework where the aforementioned ones are combined between them (Kim et al., 2021) or with other methodologies like SSL (Gao et al., 2020).

2.1.3 *Initial selection methods*

Some of the first selection criteria have come from both density-based and uncertainty-based methods. Both could produce immediate decisions over the instances so that they can also act in the stream-based scenario.

When referring to uncertainty sampling, there are simple and effective methods for classification like the *least confident* (Culotta and McCallum, 2005), *margin sampling* (Scheffer et al., 2001) and *highest entropy* (Shannon, 1948). Although these methods require the posterior probability from the classification task, for regression models, Cohn et al. (1996,?) showed that the

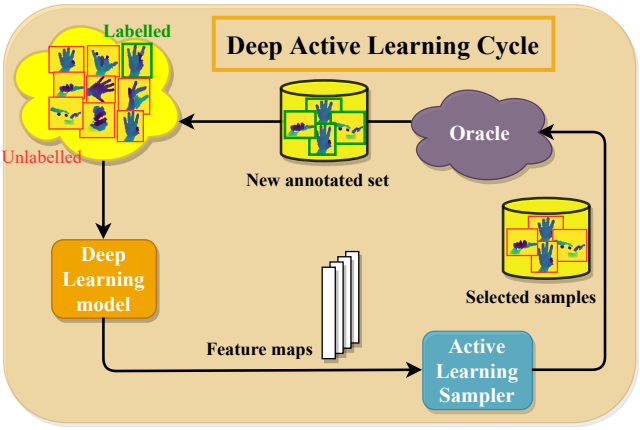

Figure 2.1: A typical DAL cycle. It starts with the initial training of labelled examples. Then, feature maps may be required by the AL sampler from the model. Once selected, the new batch is sent to the oracle for labelling. Finally, the training set is updated for another cycle.

uncertainty can be measured through variances of the estimated distribution output.

Despite the lack of popularity compared to the uncertainty ones, there have been a few initial selection methods from the representativeness category. These have benefited from the QBC structures of Naïve Bayes (McCallum and Nigam, 1998) or nearest-neighbour learners (Fujii et al., 1998) while identifying similarities in the data space. Also, one of the first notable methods to use clustering for avoiding outliers and for propagating information was proposed later on by Nguyen and Smeulders (2004).

2.1.4 *Deep Active Learning*

Since 2015, AL have started to be revisited for DL models like CNNs. Given the nature of these models, this has been addressed under the pool-based scenario. Moreover, the need for AL has come with the high demand for data that CNNs require. This new direction of research has been commonly referred to as DAL. A comprehensive survey of the most notable works has been gathered in Ren et al. (2020). A typical DAL pipeline aims to achieve a targeted accuracy with real new data in a cost-effective manner. Figure 2.1 illustrates a common cycle of the DAL methods.

This thesis founds its methodologies within DAL. Therefore, throughout the chapters, relevant literature progress will be presented of this research.

2.2 THEORETICAL FOUNDATIONS OF THE PROPOSED ACTIVE LEARNING METHODS

Each of the following subsections will present the background theory behind the AL methodologies presented in this thesis. The succession follows the order of the methods in the next chapters.

2.2.1 *Uncertainties for vision applications*

In Chapter 3, the AL proposal identifies two types of uncertainty that can interact with a DL model for 3D-HPE. Kendall and Gal (2017) have been elaborating these uncertainties before within the CV field. Thus, this subsection will highlight the theoretical findings.

To extract the two uncertainties, a Bayesian model is constructed. Such a model assumes a prior distribution $\mathcal{N}(0, I)$ over all its parameters Θ. For a set of inputs and outputs $(\mathbf{X}, \mathbf{Y})$, the Bayesian model consists of a regression function $\mathbf{f}$, specific to the 3D-HPE task. The posterior over the parameters can be expressed with Bayesian inference as $p(\Theta|\mathbf{X}, \mathbf{Y})$. Then, under a noise variation σ, the likelihood is represented as a Gaussian $p(\mathbf{y}|\mathbf{f}^{\Theta}(\mathbf{x})) = \mathcal{N}(\mathbf{f}^{\Theta}(\mathbf{x}), \sigma^2)$

Although the formulation of the Bayesian DL model might be straightforward, the challenge still stays in finding the marginal probability $p(\mathbf{Y}|\mathbf{X})$ during inference. This is essential in evaluating the posterior. In Gal and Ghahramani (2015), a Bernoulli approximation with Dropout variational inference is proposed to shift the averaging over the intractable parameters to an optimisation of a simple distribution. Therefore, the posterior is fitted with a Kullback-Leibler (KL) divergence minimisation over $q_\eta(\Theta)$, the simple distribution. The variational inference is mimicked for the Bayesian approximation by adding Dropout layers after every layer with parameters. Performing stochastic passes at testing time, with activated Dropout, is similar to sampling from the approximated posterior. This minimisation has been identified in Jordan et al. (1999) from a mixture of two Gaussians with a zero mean to one of them:

$$\mathcal{L}(\eta, p) = -\frac{1}{K}\sum_{i=1}^{K} \log p(\mathbf{y}_i|\mathbf{f}^{\tilde{\Theta}}(\mathbf{x}_i)) + \frac{1-p}{2K}\|\eta\|^2, \tag{2.1}$$

with K number of keypoints, $\tilde{\Theta}$ approximated parameters, dropout probability p and η parameters of the simple distribution. The log-likelihood for a mean square error regression loss can be written as:

$$-\log p(\mathbf{y}_i|\mathbf{f}^{\tilde{\Theta}}(\mathbf{x}_i)) \sim \frac{1}{2\sigma^2}\|\mathbf{y}_i - \mathbf{f}^{\tilde{\Theta}}(\mathbf{x}_i)\|^2 + \frac{1}{2}\log\sigma^2, \tag{2.2}$$

where σ captures the outputted noise. Conversely, this predictive variance expresses the amount of noise inherited in the data.

If the observation noise σ is a function with every input $\mathbf{x}$ (Nix and Weigend, 1994), instead of a constant, then it can be considered as an *aleatoric uncertainty* $\sigma_{al}(\mathbf{x})$, a data-dependent noise variance. The mean square loss can be rewritten as:

$$\mathcal{L}_{regression}(\mathbf{x},\mathbf{y}) = -\frac{1}{K}\sum_{i=1}^{K}\frac{1}{2\sigma_{al}(\mathbf{x})^2}\|\mathbf{y}_i - \mathbf{f}(\mathbf{x}_i)\|^2 + \frac{1}{2}\log\sigma_{al}(\mathbf{x})^2, \tag{2.3}$$

This uncertainty, however, is not captured from the variational inference to the model. As a matter of fact, the aleatoric uncertainty, in practice, is learnt with the model as a separate parameter due to its attenuation in the loss function.

To observe the noise behaviour of the model, the same Bayesian modelling is applied with Dropout sampling during testing. If M variational inferences are considered, an *epistemic uncertainty* σ_{ep}, a model-dependent variance, can be approximated from the predicted posterior $\hat{\mathbf{y}}$:

$$\sigma_{ep}^2(\mathbf{y}) \approx \frac{1}{M}\sum_{m=1}^{M}\hat{\mathbf{y}}_m^2 - \left(\frac{1}{M}\sum_{m=1}^{M}\hat{\mathbf{y}}_m\right)^2. \tag{2.4}$$

In this manner, both uncertainties can be integrated with a Bayesian approximation of a DL model, with an invasive structural change of Dropout layers. Further elaboration for 3D-HPE is presented in section 3.3.

2.2.2 *Graph Convolutional Networks*

The AL methodology from Chapter 4 is hardly linked with one of the most recent DL architectural proposals GCN. Although this architecture is addressed to assist the sampling of unlabelled data, in this section, a fundamental understanding is discussed.

Over the years, *structured graphs* $\mathcal{G}$ have been an easy and convenient way of representing the data. The instances from the data space are assigned

as *nodes* $\mathcal{V}$ while the relationships between them are represented as *edges* $\mathcal{E}$. Furthermore, the weighting between these connections can be stored in an *adjacency matrix*. The nature of updating the graphs stands on whether the propagation of the nodes is directed or undirected.

The GCN proposed by Kipf and Welling (2017) is an efficient and simplified version of spectral graph convolutions. The spectral graph theory (Rowlinson, 1996) was initially designed for eigenvalues and eigenvectors, but more recently, this has been investigated for the CNNs features in (Bruna et al., 2014). As presented by Kipf and Welling (2017), the graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ takes feature representations as nodes and their descriptive connections in an adjacency matrix A. The transition function from one convolutional layer to another can be expressed as a non-linear function:

$$f(H_l, A) = \sigma(AH_lW_l), \tag{2.5}$$

where H_l is the output of the previous layer, W_l is the matrix of weights for the l layer and σ a non-linear activation function.

When constructing the adjacency matrix for the feature nodes, computationally, it is inefficient because of two aspects. The first is that for each node, the multiplication is with its neighbours but not with itself. Therefore, a solution to overcome this is simply adding an identity matrix to the computed adjacency one $\hat{A} = A + I$. The second major flaw that commonly creates issues during training is the lack of normalisation. In Kipf and Welling (2017), a symmetric normalisation redefines the propagation rule by multiplying with the degree matrix D of the adjacency matrix:

$$f(H_l, A) = \sigma(D^{-\frac{1}{2}}\hat{A}D^{-\frac{1}{2}}H_lW_l). \tag{2.6}$$

Following these graph definitions, the GCN can be formed by stacking convolutional layers intertwined with the propagation rule.

One effect that was observed as a result of deep GCN (more than two layers) is over-smoothing (Zhao and Akoglu, 2020). The message passing spreads long-range dependencies between the node features, and then the next layers continue to smooth by centring the initial discriminative representations even more. For the datasets analysed, CORA, CiteSeer, and PubMed, it was shown in Chen, Lin, Li, Li, Zhou and Sun (2020) that two is the optimal number of layers. When measuring the mean average distance, the performance would drop for GCN configurations of 3 or more layers.

2.2.3 *Self-supervised learning*

In Chapter 5, SSL is at the root of the proposed AL framework. Since its relevance to the method, this section presents its concepts, directions and achievements. SSL, initially considered as pretext learning, is a relatively new ML direction that relies on supervisory signals to learn properties of the unobserved input data during training. Therefore, it is seen as a subset of unsupervised learning where annotations are not available. However, compared to unsupervised learning, the aim of SSL is to facilitate with pretext understanding the downstream task.

The first concepts and applications of SSL have been noticed in the NLP domain, where there is a large dictionary of words. Research works like Joulin et al. (2016); Mikolov et al. (2013); Pennington et al. (2014) have greatly benefited in building up the corpus of words with SSL. It is just recently that the application domain expanded to CV. This was possible by associating the SSL as an Energy-based function (EBF).

In principle, an EBF tracks the compatibility between an observed input and its proposed prediction. For small energy, there is high compatibility, while it is expected high energy in case of incompatibility. Two directions have been considered to model the EBF: joint embedding and contrastive learning. Generally, the joint embedding maps the EBF through two almost identical networks. The input and its proposed prediction are passed separately, and their outputted vectors are used to compute the energy. A disadvantage of this approach is the mode collapse when the joint pipeline produces identical output embeddings during training. In an attempt to deal with mode collapse, joint embedding adds regularisation techniques. For solutions like BYOL (Grill et al., 2020), an asymmetric pipeline is presented together with a strong variation of data augmentation. SimSiam (Chen and He, 2021) or SwAV (Caron et al., 2020), on the other hand, make use of virtual target embeddings coming from similar groups.

Another way of dealing with mode collapse is contrastive learning, suggested as the second direction of SSL. Here, the EBF is expecting pairs of positive and negative inputs. In the majority of cases, negative inputs are considered far in the features space, while positive examples are likely augmented versions of the observed input. In proposals like SimCLR (Chen, Kornblith, Norouzi and Hinton, 2020) or MoCo (He et al., 2020), a noise-contrastive estimation, InfoNCE (Oord et al., 2018), is optimised as a loss function. Moreover, integrating dictionaries in MoCo and MoBY (Xie et al., 2021) to compute the loss over time has helped to improve the visual representations.

In terms of achievements, SSL is evaluated as an up-and-coming area of research. The visual concepts learnt with SSL have started to produce high accuracy for tasks like image classification, object detection or image segmentation (Xie et al., 2021). In some cases, it has even over-passed the full supervised training (Goyal et al., 2021), making it very efficient in handling large-scale datasets.

2.2.4 *Variational auto-encoder and Product of Experts*

The methodology from the final chapter 6 is inspired by the cross-modality solutions introduced to 3D-HPE by Cross-Modal Spurr et al. (2018), and AlignMM Yang et al. (2019). Both works use generative algorithms with frameworks based on the VAE architectures. These solutions tackle multiple modalities and utilise the domain translation between them to achieve state-of-the-art results. This further motivated the AL approach in sampling for 3D-HPE. Thus, in this part, it is supportive to briefly detail the VAE model and further delve into the methodology of AlignMM.

The VAE, part of DL generative models, aims to reduce the dimensions of the feature representation by reconstructing the input images. The VAE, architecturally, is composed of neural networks with an encoder and a decoder design. Compared to auto-encoders, the VAE adds regularisation to avoid over-fitting and maintain a good generative latent space. Therefore, the inputs are encoded as a distribution over the latent space and sampled points from the latent distribution are decoded to compute the reconstruction error. The regularisation term is done with variational inference, similar to the Bayesian approximation. A KL divergence is computed with the reconstruction loss by aligning to a normal distribution the latent mean and variance.

AlignMM is an extension of the Cross-Modal generative method for multi-modal 3D-HPE. Considering $\mathbf{x}$ from some modality and $\mathbf{y}$ the output from a target modality, the multi-modal VAE estimates it by maximising the evidence lower bound (ELBO). If $\mathbf{z}$ is the latent variable, then the ELBO is:

$$\mathcal{L}_{ELBO}(\mathbf{x};\mathbf{y};\theta,\phi) = E_{z\sim q_\phi} \log p_\theta(\mathbf{y}|\mathbf{z}) - \beta KLD(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})), \tag{2.7}$$

where KLD is the KL divergence, β is the disentangling factor from Higgins et al. (2017), $p(\mathbf{z})$ the Gaussian prior and the encoder - decoder probabilities $q_\phi(\mathbf{z}|\mathbf{x}), p_\theta(\mathbf{y}|\mathbf{z})$. When adding a second modality $\mathbf{w}$, the ELBO can be extended as:

$$\begin{aligned}\mathcal{L}_{\mathrm{ELBO}}(\mathbf{x},\mathbf{w};\mathbf{y},\mathbf{w};\theta_{\mathbf{y}},\theta_{\mathbf{w}},\phi_{\mathbf{x},\mathbf{w}}) = \mathrm{E}_{z\sim q_{\phi_{\mathbf{x},\mathbf{w}}}} \log p_{\theta_{\mathbf{y}}}(\mathbf{y}|\mathbf{z}) + \\ \lambda_{\mathbf{w}} \mathrm{E}_{\mathbf{z}\sim q_{\phi_{\mathbf{x},\mathbf{w}}}} \log p_{\theta_{\mathbf{w}}}(\mathbf{w}|\mathbf{z}) - \\ \beta \mathrm{KLD}(q_{\phi_{\mathbf{x},\mathbf{w}}}(\mathbf{z}|\mathbf{x},\mathbf{w}) \| p(\mathbf{z})),\end{aligned} \tag{2.8}$$

with $\lambda_{\mathbf{w}}$ regulating parameter for the reconstruction from $\mathbf{w}$ to $\mathbf{y}$. If the encoder takes inputs only from $\mathbf{x}$, this can be simplified to $\mathcal{L}_{\mathrm{ELBO}}(\mathbf{x};\mathbf{y},\mathbf{w};\theta_{\mathbf{y}},\theta_{\mathbf{w}},\phi_{\mathbf{x}})$.

In order the align the two latent spaces, another KL divergence is minimised between $q_{\phi_{\mathbf{x},\mathbf{w}}}$ and $q_{\phi_{\mathbf{x}}}$. The new combined loss for the $\mathbf{z}_{\mathrm{joint}}$ joint latent space is:

$$\begin{aligned}\mathcal{L}(\phi_{\mathbf{x},\mathbf{w}},\phi_{\mathbf{x}},\theta_{\mathbf{y}},\theta_{\mathbf{w}}) = \mathcal{L}_{\mathrm{ELBO}}(\mathbf{x},\mathbf{w};\mathbf{y},\mathbf{w};\theta_{\mathbf{y}},\theta_{\mathbf{w}},\phi_{\mathbf{x},\mathbf{w}}) + \\ \mathcal{L}_{\mathrm{ELBO}}(\mathbf{x};\mathbf{y},\mathbf{w};\theta_{\mathbf{y}},\theta_{\mathbf{w}},\phi_{\mathbf{x}}) - \\ \beta_{\mathrm{joint}} \mathrm{KLD}(q_{\phi_{\mathbf{x},\mathbf{w}}}(\mathbf{z}_{\mathrm{joint}}|\mathbf{x},\mathbf{w}) \| q_{\phi_{\mathbf{x}}}(\mathbf{z}_{\mathbf{x}}|\mathbf{x}).\end{aligned} \tag{2.9}$$

The embedding in the same space for $\mathbf{z}_{\mathrm{joint}}$ and $\mathbf{z}_{\mathrm{x}}$ might be suitable for two modalities, but if they increase, the joint encoder $q_{\phi_{\mathbf{x},\mathbf{w}}}$ is challenging to learn. As a solution, Product of Experts (PoE) () alternative is proposed for aligning the latent spaces. The PoE comprises of the two Gaussian distributions $q(\mathbf{z}|\mathbf{x})$ and $q(\mathbf{z}|\mathbf{w})$. The final objective function with the shared decoders can be rewritten as:

$$\begin{aligned}\mathcal{L}(\phi_{\mathbf{w}},\phi_{\mathbf{x}},\theta_{\mathbf{y}},\theta_{\mathbf{w}}) = \mathcal{L}_{\mathrm{ELBO}}(\mathbf{x},\mathbf{w};\mathbf{y},\mathbf{w};\theta_{\mathbf{y}},\theta_{\mathbf{w}},\phi_{\mathbf{x}},\phi_{\mathbf{w}}) + \\ \mathcal{L}_{\mathrm{ELBO}}(\mathbf{x};\mathbf{y},\mathbf{w};\theta_{\mathbf{y}},\theta_{\mathbf{w}},\phi_{\mathbf{x}}) + \\ \mathcal{L}_{\mathrm{ELBO}}(\mathbf{w};\mathbf{y},\mathbf{w};\theta_{\mathbf{y}},\theta_{\mathbf{w}},\phi_{\mathbf{w}})\end{aligned} \tag{2.10}$$

Explicitly, the joint objective with the product of Gaussian experts $G^*(\mathbf{z}_{\mathbf{x}},\mathbf{z}_{\mathbf{w}})$ is:

$$\begin{aligned}\mathcal{L}(\phi_{\mathbf{w}},\phi_{\mathbf{x}},\theta_{\mathbf{y}},\theta_{\mathbf{w}}) = \mathrm{E}_{\mathbf{z}_{\mathrm{joint}}\sim G^*(\mathbf{z}_{\mathbf{x}},\mathbf{z}_{\mathbf{w}})} \log p_{\theta}(\mathbf{y},\mathbf{w}|\mathbf{z}_{\mathrm{joint}}) + \\ \mathrm{E}_{\mathbf{z}_{\mathbf{x}}\sim q_{\phi_{\mathbf{x}}}} \log p_{\theta}(\mathbf{y},\mathbf{w}|\mathbf{z}_{\mathbf{x}}) \mathrm{E}_{\mathbf{z}_{\mathbf{w}}\sim q_{\phi_{\mathbf{w}}}} \log p_{\theta}(\mathbf{y},\mathbf{w}|\mathbf{z}_{\mathbf{w}}) - \\ \beta(\mathrm{KLD}(q_{\phi}(\mathbf{z}_{\mathbf{x}}|\mathbf{x}) \| p(\mathbf{z})) + \mathrm{KLD}(q_{\phi}(\mathbf{z}_{\mathbf{w}}|\mathbf{w}) \| p(\mathbf{z}))).\end{aligned} \tag{2.11}$$

In case of more than two modalities, other KL divergences would be added to force all the latent spaces into a joint one. Apart from the joint reconstruction loss, the authors of AlignMM integrated dedicated losses and tricks for

the 3D-HPE task. However, this defies the purpose of the multi-modal AL sampler.

2.3 3D HAND POSE ESTIMATION

2.3.1 *Definition*

Given a detected area of a hand from an input space, the 3D-HPE task is defined as the process of mapping and localising in 3D coordinates the shape and structure of the hand with a ML algorithm. For a clearer explanation of the mapping process, a skeleton model of the hand is initially proposed with all the kinematic parameters. Most of the recent state-of-the-art solutions that come from DL have adopted the skeleton model studied in Erol et al. (2007).

This kinematic hand model is formed by following the 21 skeleton joints that have some DOF. These joints follow the medical name location and correspondence to each finger. Thus, six joints are describing the palm, one root at the carpals and five metacarpophalangeal (MCP), and three more joint locations corresponding to each finger: proximalcarpophalangeal (PIP), distalcarpophalangeal (DIP) and tip.

Breaking down to the 3D coordinates for each joint, these are calculated by referencing the location to the recording device. The depth information can be extracted directly from the acquisition system in the case of depth camera sensors. However, this is not the case when using RGB cameras, where the depth information must be modelled differently (stereoscopic cameras, time-of-flight sensors etc.).

Moving into the challenges of 3D-HPE, several have been identified:

- occlusions, including self-occlusion - the camera projections will lack of meaningful information when overlapping with joints occurs;
- real-time processing for each frame - a system to run the ML model fast enough;
- lighting variation - typical issue for RGB inputs;
- the last, but not the least, mapping highly dimensional non-linear functions.

DL models and datasets from better-capturing systems have addressed these issues to a great extent.

2.3.2 *Categories*

There are several criteria that 3D-HPE models can be classified. With the great availability of sensors and depth cameras, starting with the Microsoft Kinect from 2011, depth-based models (Oberweger et al., 2015; Ye et al., 2016) have become the echelon due to their robustness to noise and illumination variance. Previously, the datasets and models tackled the RGB environment primarily (Ge et al., 2019; Tang et al., 2015; Zimmermann and Brox, 2017). More recently, the 3D-HPE task benefits from multi-camera systems with one or two modalities, RGB-D (Simon et al., 2017). Therefore, a clear distinction between these models stands on the input space type.

Following the taxonomy from Erol et al. (2007), the 3D-HPE solutions can be classified according to the minimisation objective into *generative* (detection-based) (Baek et al., 2018*b*; Chang et al., 2018; Tang et al., 2015) , *discriminative* (regression-based) (Chen, Wang, Guo and Zhang, 2020; Oberweger and Lepetit, 2017; Oberweger et al., 2015; Rad et al., 2018) and hybrid combination of both (Baek et al., 2019; Ge et al., 2019; Sanchez-Riera et al., 2018; Yang et al., 2019; Ye et al., 2016). In principle, the generative models aim to minimise generated poses through regions or heat maps, while the discriminative ones direct the objective through the regression of features or key points.

Depending on their architectural design, these DL models can also be categorised as *hierarchical, structured, multi-stage, 2D/3D* and *residual*. Hierarchical models split the estimation of the hand into multiple parts, either on joints or on fingers and palm (Guo et al., 2017; Ye et al., 2016). Multi-stage and structured methods are conceptually similar as both aim to propagate initial guidance for estimation. The key difference is that structured models (Chen, Wang, Guo and Zhang, 2020; Oberweger and Lepetit, 2017; Oberweger et al., 2015; Tang et al., 2014; Xiong et al., 2019) rely initially on hand kinematic constraints. In contrast, the multi-stage ones (Baek et al., 2019, 2018*b*; Guo et al., 2017) add similar constraints or specific modules to improve the first hand estimation. The architecture for the residual methods is founded on the ResNet (He et al., 2016) CNNs. This has been one of the best solutions in capturing visual representation due to the residual blocks. Models that have adopted this architecture are Baek et al. (2019); Chang et al. (2018); Chen, Wang, Guo and Zhang (2020). Finally, the 2D/3D category designs the DL model either with 2D (Oberweger et al., 2015; Xiong et al., 2019; Ye et al., 2016) or volumetric CNNs (Chang et al., 2018; Huang et al., 2018).

2.3.3 *Deep Learning solutions*

Multiple DL solutions for 3D-HPE have existed throughout the past years. On their achievements, the mean joint error performance has decreased to 7 mm for depth-based datasets like NYU (Tompson et al., 2014) or ICVL (Tang et al., 2014).

At the time of developing the AL algorithm from Chapter 3, the hierarchical spatial attention with Particle Swarm Optimisation (PSO) proposed by Ye et al. (2016) was one of the top methods. Therefore, this has motivated an initial deployment of it as the learner to the AL pipeline. Further on, due to its architectural complexity, the method from Chapter 3 approached a more simple design from a previous state-of-the-art, DeepPrior (Oberweger et al., 2015). In this section, the mentioned hand pose estimators are presented, together with the current state-of-the-art volumetric residual network V2VPoseNet (Chang et al., 2018).

The work of Ye et al. (2016) proposes a hierarchical model by starting from a network that produces feature maps for the palm area and by continuing with two cascaded networks for each finger. One of the cascaded networks identifies the PIP joints and the other the DIP and the fingertip. The training starts with the palm network, and the other ten CNNs are trained sequentially. Moreover, spatial attention modules that take the previous feature maps, rotation and space transform are added between the cascaded networks to achieve viewpoint invariance. For the final refinement, partial PSO is applied at each network to maintain the kinematic constraints. The method adopts a regression-based minimisation where all the networks output the 3D coordinates to their targeted joint(s).

One of the first works to achieve state-of-the-art in 3D-HPE with CNNs is DeepPrior (Oberweger et al., 2015). It consists of two parts: the first contribution is by investigating several feature extractors, including a hierarchical one as in Ye et al. (2016); the second stands on introducing a bottleneck for hand constraints before the last regression layer. The bottleneck is initialised with Principal Component Analysis (PCA) of hand poses. The design of the CNN feature encoders is a classic combination of one to three convolution layers and max pooling. For the objective, DeepPrior also follows a discriminative minimisation loss to the 3D joint coordinates.

V2VPoseNet (Chang et al., 2018) has been the consistent state-of-the-art solution from benchmarks like ICVL, NYU, MSRC (Sharp et al., 2015) to challenges like Hands in a Million 2017 (Yuan, Ye, Garcia-Hernando and Kim, 2017*a*). Its success mainly comes from breaking the highly non-linear

regression to a voxelised space and applying 3D (volumetric) residual CNNs. Therefore, the first step is to create from the 2D depth image the 3D map of hand voxels. This map will set up the boundaries of the hand location. The second step consists of a 3D stacked hourglass architecture (Newell et al., 2016) that generates 3D heat maps with joint location likelihoods.

The 3D voxelised space is created with a pre-trained DeepPrior++ network (Oberweger and Lepetit, 2017). This provides a good reference point to threshold the 3D areas of the hand. V2VPoseNet transforms the hourglass architecture with 3D convolutions, pooling and up-sampling layers. To this extent, the architecture encloses a large number of parameters, but the key is to keep the 3D voxelised space relatively small (at 44x44x44).

2.4 DATASETS FOR HAND POSE ESTIMATION

2.4.1 *Dataset Engineering*

Dataset engineering is a particular essential field required by ML scientists. Without a robust and elaborated collection of data, the ML algorithms may lie just in a concept or hypothetical phase. As described in the introduction, dataset engineering generally assumes four stages: *planning*, *acquisition*, *annotation* and *evaluation*.

Depending on the planning, the data acquisition may include steps such as generation and augmentation, especially if real data is unavailable or a synthetic environment is proposed. Furthermore, the data acquisition may be refined with data cleaning or optimisation tools either to discard noisy data or to integrate an AL selection pipeline.

The annotation process might be even more challenging than the acquisition stage, however, in the planning phase, automatic systems may come in hand. Outsourcing to third parties like Amazon Mechanical Turk has also been a frequent option for large-scale datasets with certain risks like confidentiality and labelling quality. The next sections will detail the data engineering process for the 3D-HPE datasets used in this thesis.

2.4.2 *Depth image datasets*

The depth-based 3D-HPE benchmarks that were commonly evaluated at the start of this thesis are ICVL (Tang et al., 2014), BigHand2.2M (Yuan, Ye, Stenger, Jain and Kim, 2017) and NYU (Tompson et al., 2014). Therefore, in

the next chapter, these benchmarks are tackled with AL selection techniques for the first time.

ICVL has been one of the first milestones for depth-based hand datasets. In this regard, it has the lowest number of frames between the three of them, totalling 17,604. Despite this, the annotation is done effectively with 3D tracking (Melax et al., 2013) and a second manual refinement. This, however, yields approximate coordinates, and, as a result, it trains the models with unnatural poses. Another drawback of ICVL is the limited articulation and viewpoint space. During the acquisition process, the ten subjects are recorded with Intel's Creative interactive gesture camera.

Published in 2014, the *NYU* hand dataset comprises 72,757 training images and a testing set of 8,252. Although the dataset is gathered from only two subjects, the hand poses are recorded with 3 Microsoft Kinect cameras with three viewpoints (frontal, left and right side). Compared to ICVL, the images have double the resolution at 640x480, and they are available in RGB-D format. The annotation follows similar stages to ICVL of hand tracking and a PSO refinement. Another advantage of NYU consists in providing more accurate annotations with one of the first CNN that takes multi-resolution inputs and generates heat maps for pose recovery.

The first real large-scale hand dataset with an automatic annotation system is *BigHand2.2M*. It was published in 2016, and since then, it has been integrated into several hand pose challenges like "Hand in a Million 2017" (Yuan, Ye, Garcia-Hernando and Kim, 2017*b*). During the acquisition stage, ten subjects were recorded with complete coverage of camera viewpoints and a wide range of articulations. The camera in BigHand2.2M is Intel's RealSense SR300 with the same resolution as NYU. The annotation stage is somehow combined with the acquisition process. This is because the subjects have attached to their hands six 6D magnetic sensors that are further connected to a trakSTAR tracking system (Yuan, Ye, Stenger, Jain and Kim, 2017). As it is stated in the name, BigHand2.2M consists of 220,000 images for each subject. In terms of training and testing, the authors proposed a split of 9 subjects for training and one for testing.

2.4.3 *Multi-modal datasets*

In Chapter 6, the AL methodology investigates multi-modality for 3D-HPE. To fit this goal, three well-known RGB-D datasets are approached: STB (Zhang et al., 2017), RHD (Zimmermann and Brox, 2017) and FreiHand (Zimmermann

et al., 2019). These datasets have at least the colour and depth modality so that the learner can simulate the AL training and sampling.

Moving on the *STB dataset*, it contains 18,000 stereo image pairs with frontal camera viewpoint. Each hand articulation is framed in six different background scenarios with 640x480 pixels. The articulation space is also limited in 12 sequences with 1,500 frames each. For the annotation process, however, it is done manually with the help of the stereo and depth cameras setup. Part of the dataset contains post-processed segmentation images. This can be used either for a hand segmentation task or as a separate modality.

While the STB dataset presents real hand images, *RHD* provides rendered synthetic data from 20 subjects performing 39 different actions. This allows them to generate 41,258 and 2,728 training and testing images with 320x320 resolution. Regarding annotation, RHD, as STB, also creates segmentation maps, detailing them to palms and fingers. So far, it has not been shared the process of finding the 21 key points. An assumption is that the generator already contains the hand models, and the corresponding joint locations are referenced to the 3D synthetic studio.

The last multi-modal 3D-HPE dataset tackled in this thesis is *FreiHand*. Compared to the previous two datasets, it is one of the most recent, and it is the single one to include object interaction and occlusion. In addition, FreiHand gathers 130,240 training images from four camera viewpoints and 3,960 images to evaluate. Although depth maps are not available, any of the four viewpoints can be seen as a separate modality. Moreover, FreiHand can provide the 224x224 RGB images, their segmentation mask and the synthetic hand model of MANO (Romero et al., 2017). The annotation process is semi-automatic by fitting the MANO model in the multi-view system. The fitting network also produces confidence scores so that an oracle can manually assist the annotation process.

3

ACTIVE LEARNING FOR BAYESIAN 3D HAND POSE ESTIMATION

3.1 COURSE OF OBJECTIVES WITH JUSTIFICATION

This chapter focuses on answering the first problem from the Introduction:

> "Given a DL-based 3D hand pose estimator, how can the learner reach minimal error with AL when meaningful real sampled data are provided?"

Firstly, the question arises from a lack of exploratory research of AL for 3D-HPE. This initial curiosity is even more sustained by the recent data-demanding DL algorithms.

Secondly, as mentioned in section 1.1, with the advancements of the DL models, a gap has been gradually set up between the models and the representativeness of the datasets. This comes after, at the time of publishing the largest depth-based dataset, BigHand2.2M (Yuan, Ye, Stenger, Jain and Kim, 2017), it was shown that models suffer little to no precision when trained with fractions of the entire dataset. Even though DL models are known to improve with more data, this process becomes extremely inefficient when data annotation is commonly noisy, tedious and expensive.

Finally, AL is proposed to optimise the data sampling, but part of the question is to know how. Before addressing this, it is important first to understand the task by deploying a standard DL architecture. In the majority of cases, the 3D-HPE is presented in a discriminative manner with a final regression to the joint coordinates. One main disadvantage is that the predicted posterior does not provide any confidence information with its estimation. Therefore, in the first part, a Bayesian approximation is presented for the 3D-HPE learner, similarly to Kendall and Gal (2017), so that two types of uncertainties can be modelled. The two uncertainties, aleatoric and epistemic, represent the variances of the data-dependent and model-dependent samples, respectively.

This Bayesian adaptation also facilitates the investigation and proposals of AL selection function. Because it is the first approach for AL, classification-based selection functions are not transferable to the regression task. One of the other ways of sampling data apart from random sampling is with

the state-of-the-art geometric function CoreSet (Sener and Savarese, 2018). CoreSet uses the latent space distribution to sample images far from the annotated centroids in Euclidean space. The novel method presented in this chapter, **CKE**, extends the limitations of CoreSet in calculating the distances according to the epistemic variances. CKE combines in this manner the representativeness of the data together with the level of uncertainty that each data impacts the space. Furthermore, the learner integrates the aleatoric uncertainties in its loss function to attenuate the influence of difficult or noisy samples.

In a first attempt of modelling the aleatoric and epistemic uncertainties, the learner deployed was following the hierarchical spatial attention CNN from Ye et al. (2016). However, due to the training complexity of evaluating 11 networks with each AL cycle, the architecture of the learner changed to a single deep CNN as in DeepPrior (Oberweger et al., 2015). This simplifies the experimental research in finding the optimal AL selection strategy. In Appendix A, an extension of the two uncertainties concerning the initial learner supports the transferable insights to DeepPrior.

In summary, the next sections go through a literature review 3.2, the CKE methodology with the Bayesian approximation 3.3 and culminate with a thorough set of experiments for depth-based 3D-HPE datasets 3.4.

3.2 RELEVANT LITERATURE

3.2.1 *3D Hand Pose Estimation*

The first comprehensive review in hand pose estimation that established the current taxonomy is published in Erol et al. (2007). Together with the deep learning popularity and easy access to depth camera sensors, 3D-HPE has gained a deep interest in the computer vision community. Earlier state-of-the-art methods favoured combination of discriminative (random forest (Tang et al., 2014) or CNN-based(Oberweger et al., 2015)) and generative solutions Baek et al. (2018*b*) as in (Oberweger and Lepetit, 2017; Ye et al., 2016). Due to the past concepts from Charles et al. (2017), volumetric representations as V2VPoseNet (Chang et al., 2018) managed to out-stand by compensating for the high non-linearity in direct regressions. These recent works (Chang et al., 2018; Wan et al., 2018; Xiong et al., 2019) have obtained impressive results with average 3D joint errors below 10 mm on datasets like NYU (Tompson et al., 2014), ICVL (Tang et al., 2014) or BigHand2.2 (Yuan, Ye, Stenger, Jain and Kim, 2017). While tackling uncertainty exploration and

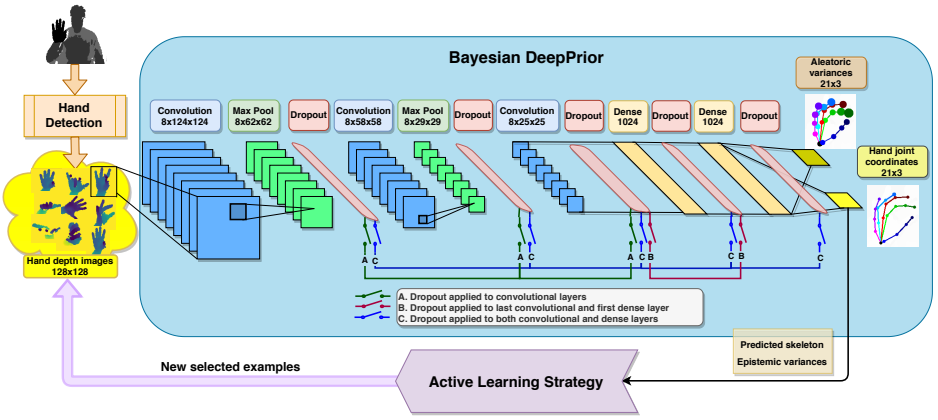

Figure 3.1: Schematic diagram showing the end-to-end pipeline of the AL method. The depth image of the hand is pre-processed and fed to the Bayesian DeepPrior Network. This network is trained to minimise the objective function given in Equation 3.2. The aleatoric and epistemic uncertainties of a sample that we obtain from M number of Monte Carlo Dropouts are passed on to the selection criteria for querying unlabelled examples. The selection method is highlighted in the Algorithm 1. The switches in the Figure are present to indicate the activation of Dropout in different Bayesian approximations.

data representatives of the model, for simplicity and efficient analysis, the standard DeepPriorOberweger et al. (2015) architecture is deployed. Even though the accuracy is lower than current arts due to relatively lesser model parameters, the AL method is generic, thus, the insights obtained can be easily transferred to deeper and generative models.

3.2.2 *Bayesian Deep Learning*

Recently, there have been several investigations in quantifying and representing uncertainty in CNNs. A novel approach is to approximate variational inference in a Bayesian implementation (Gal and Ghahramani, 2015) for image classification (Gal and Ghahramani, 2016). Furthermore, the types of uncertainties (aleatoric or epistemic) and the ways to employ these uncertainties in both regression and classification tasks are presented in Kendall and Gal (2017); Kendall et al. (2018). The novel approximation of Bayesian (Gal and Ghahramani, 2016) for deep learning by using the Dropout layers(Srivastava et al., 2014) reduced the computational complexities of the naïve Bayesian Neural Network implementation. An analysis of uncertainty for the active learning framework is presented in Zeng et al. (2018). However, the eval-

uation is conducted on a small-scale dataset such as MNIST (LeCun and Cortes, 2010) classification. In contrast, the approach in this chapter analyses a more difficult and large-scale experimental setup.

Another recent work on image classification Beluch Bcai et al. (2018) proposed ensemble-based active learning and demonstrated outperforming the Dropout Bayesian approximationGal and Ghahramani (2016). Again, the experiments are constrained on small-scale scenarios. On these premises, the methodology analyses the exploration of both data and model-dependent uncertainties similar to Kendall and Gal (2017), but for the large-scale and more challenging problem i.e. 3D-HPE. Furthermore, this will integrate the risk measurement concerns presented in Osband (2016) through the aleatoric uncertainty.

3.2.3 *Active Learning Methods*

This branch of machine learning was explored in the need for informative datasets that are, in most cases, model-dependent. With the advances in deep neural networks, the research community began to approach the classical approaches despite the lack of integration in online training (an essential part of the active learner methodology). The most common scenario used, pool-based sampling, limits the live model refinement by performing offline training Gal et al. (2017); Sener and Savarese (2018); Sinha et al. (2019); Yoo and Kweon (2019). On the other hand, pool-based active learning opened a new direction of research for deep neural networks (Gal and Ghahramani, 2016). This has evaluated with what minimum percentage of the training set the model can achieve the same accuracy as the entire one. A theoretical approach has been explored in Buyu and Vittorio (2017) over the human pose estimation problem by actively collecting unlabelled data from the heat-map output of the Convolutional Pose Machines (CPMs). However, the proposed methodology is driven independently from the output as it relies directly on the predicted hand skeleton. In terms of pose estimation, a practical application of active learning for hands has been conducted in Hsiao et al. (2016), where KD-trees are applied to guide the camera movement to a more informative viewpoint.

3.3 METHODOLOGY

This section starts with formulating the 3D-HPE problem as a Bayesian approximation inspired by Kendall et al. (2015); Kendall and Gal (2017) for semantic segmentation and depth regression. A framework (similar to Kendall et al. in Kendall and Gal (2017)) is adopted to model the two uncertainties: aleatoric and epistemic. These two statistics characterise the important but complementary aspect of the model when trained from data. In particular, *aleatoric uncertainty* captures uncertainty due to noisy training examples, which can not be eradicated from the model even if there is plenty of training examples. Whilst, *epistemic uncertainty* quantifies the ignorant aspect of the model parameters, which can be addressed with the availability of training examples. For an AL application, the aleatoric uncertainty plays a key informative role in avoiding annotating noisy samples when acquiring new data. Moreover, these uncertainties would also balance hard samples from which the current learner can't further improve. Besides, the epistemic variance helps the AL sampling method in indicating relevant unseen data for the learner. Given the parameters inferred from the Bayesian hand pose estimator in the second part, an acquisition method is investigated, commonly known as a sampler in the AL scheme, by enclosing the uncertainty variances. Hence, the contribution relies on the analysis of the Bayesian 3D-HPE approximation (comprises the learner component in AL), together with a novel selection mechanism. Figure 3.1 summarises the proposed pipeline.

3.3.1 *Bayesian 3D Hand Pose Estimator*

3.3.1.1 *3D Hand Pose Estimator*

In classification tasks, uncertainty is estimated by the posterior probability of a class. However, the 3D-HPE, a regression problem, maps hand (depth) images to 3D coordinates of the hand joint locations. The method to regress the skeleton coordinates, along with modelling the mentioned uncertainties, is described in this part. The inputs to the hand pose estimator are cropped images centred on the hand region. Reducing the potentially cluttered or noisy areas in the depth scene motivates the main advantage of this practice. For annotated ground truths, the hand skeletons are formed following Erol et al. (2007) by 21 volumetric joints. With this information, common practice to tackle the translation of different palm sizes is to set the origin to the root

palm and normalise the UV coordinates with the distance from it to the index finger MCP.

In order to topologically evaluate the regression uncertainties, a standard *DeepPrior* (Oberweger et al., 2015) architecture is designed. This comprises a convolutional feature extractor and a dense regression. Given $\mathbf{x} \in \mathbb{R}^{w \times h}$, a 2D cropped hand image with w width and h height is inferred through the CNN, together with its corresponding ground-truth $\mathbf{y} \in \mathbb{R}^{21 \times 3}$. Structurally, the feature extractor is composed of three block groups of convolution, max pool and LeakyReLU activation. The flattened output of the feature extractor is regressed in the end by 2 dense layers (Figure 3.1). For simplicity, the final PCA layer is excluded from the initial design of Oberweger et al. (2015), and it directly predicts the normalised UVD joint coordinates $\hat{\mathbf{y}}$ (description of the pipeline in Fig. 3.1). Finally, the parameters are optimised by Stochastic Gradient Descent (SGD) to minimise the mean squared error (MSE). The objective function is as below:

$$\mathcal{L}(\mathbf{x}, \mathbf{y}; \Theta) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{K} \sum_{k=1}^{K} \|\mathbf{y}_{i,k} - \hat{\mathbf{y}}_{i,k}\|^2 \right), \tag{3.1}$$

where, n is the data size, Θ represents model parameters and K is the number of joints.

3.3.1.2 *Bayesian DeepPrior*

The *Bayesian DeepPrior* is described here in detail. As stated before, unlike in the classification model, it is not straightforward to model the uncertainties in the regression model. In order to make the DeepPrior a probabilistic model, Dropout (Srivastava et al., 2014) is introduced on its layers similar to Gal and Ghahramani (2016).

A Bayesian Neural Network (BNN) sets a normal distribution $\Theta \sim \mathcal{N}(0, I)$ on its weight parameters as prior. To approximate the posterior distribution over the weights $f(\Theta|\mathbf{x}, \mathbf{y})$, a KL divergence is minimised of a variational inference distribution $q(\Theta)$ and its posterior: $KL(q(\Theta)\|f(\Theta|\mathbf{x}, \mathbf{y}))$. To minimise the KL divergence loss, Monte Carlo (MC) Dropout is applied during the variational inference, and the MSE of joint location's prediction is optimised. This minimisation is equivalent to the KL divergence approximation as detailed in Section 2.2.1. Once the Dropout is kept active over the entire Θ, the Bayesian approximation can be obtained of the posterior's mean and variance.

As stated in Osband (2016), the uncertainty variance of the BNN consists of the Bayes risk rather than a systematic uncertainty. Therefore, the outputs of the final dense layer are learnt together with the aleatoric uncertainty by balancing the MSE objective function during training as well. Specifically, an aleatoric variance is allocated for each hand joint coordinate. Hence, the new objective function to model such uncertainties is defined as in Kendall and Gal (2017):

$$\mathcal{L}_B(\mathbf{x}, \mathbf{y}; \Theta) = \frac{1}{n} \sum_{i=1}^{n} \Big(\frac{1}{K} \sum_{k=1}^{K} \frac{1}{2} e^{-\alpha_{i,k}} \|\mathbf{y}_{i,k} - \hat{\mathbf{y}}_{i,k}\|^2 + \frac{1}{2} \alpha_{i,k} \Big), \tag{3.2}$$

with the logarithmic variance $\alpha_{i,k} = \log \hat{\sigma}^2_{al}$ and $\hat{\sigma}^2_{al}$, the aleatoric variance. Thus, the learnt numerically-stable logarithmic tracks the noise present in the data.

To summarise, by applying MC Dropout, a Bayesian DeepPrior is obtained, and it generates a mean value for each joint coordinate. After variational inferences, we also evaluate its epistemic and learnt aleatoric variances. For M times passes of a sample, the epistemic uncertainty is estimated as below:

$$\hat{\sigma}^2_{ep} \approx \frac{1}{M} \sum_{m=1}^{M} \hat{\mathbf{y}}^2_m - \Big(\frac{1}{M} \sum_{m=1}^{M} \hat{\mathbf{y}}_m \Big)^2 \tag{3.3}$$

Finally, the combined variances for a predicted skeleton $\hat{\mathbf{y}}$ can be expressed as:

$$\hat{\sigma}(\hat{\mathbf{y}}) \approx \frac{1}{M} \sum_{m=1}^{M} \hat{\mathbf{y}}^2_m - \Big(\frac{1}{M} \sum_{m=1}^{M} \hat{\mathbf{y}}_m \Big)^2 + \frac{1}{M} \sum_{m=1}^{M} \hat{\sigma}^2_{al_m}. \tag{3.4}$$

In the experiments section 3.4.2, a study on the number of M vs the mean joint error stabilisation is presented.

From an architectural perspective, the Bayesian DeepPrior contains Dropout layers after every convolutional and dense layer. In practice Kendall et al. (2015); Kendall and Gal (2017), it has been shown that this model may suffer from strong regularisation. Thus, similarly to Kendall et al. (2015), different variants where Dropout is applied to designated locations are analysed. These probabilistic variants of Bayesian DeepPrior are simulated through switches in the proposed pipeline3.1 accordingly:

- **A** - only to all convolutional layers;
- **B** - centrally to the last convolutional layer and the first dense layer;
- **C** - throughout the entire DeepPrior architecture.

To fine-tune the design of the Bayesian DeepPrior architecture, a cross-validation study is performed on the NYU Hand dataset Tompson et al. (2014) in section 3.4.2.

3.3.2 *Active Learning Framework*

In this section, the AL process is described for DL together with the proposed acquisition function adapted for the Bayesian DeepPrior, CKE.

3.3.2.1 *Pool-based Active Learning Strategy*

As AL has gained stronger interest in DL, a pool-based scenario has become a standard methodology to overcome the data-greedy models with their slow training process (Settles, 2009). Therefore, the pool-based active learning considers a scenario with an initial annotated set $\mathbf{s}^0$ and an available unlabelled dataset $\mathcal{U}_{pool}$. The goal is to find the least amount of L annotated subsets $\mathbf{s}^1, \mathbf{s}^2 \dots \mathbf{s}^L \subset \mathcal{U}_{pool}$ so that the targeted mean squared joint error is achieved. Given an acquisition function $\mathcal{A}$, this can be summarised under the following equation:

$$\min_{L} \min_{\mathcal{L}_B} \mathcal{A}(\mathcal{L}_B; \mathbf{s}^1, \mathbf{s}^2 \dots \mathbf{s}^L \subset \mathcal{U}_{pool}). \tag{3.5}$$

In the next section, the function $\mathcal{A}$ is formulated to suit the Bayesian DeepPrior architecture.

3.3.2.2 *Combination of Geometric and Uncertainty Acquisition Function (CKE)*

Acquisition functions have been extensively developed for classification tasks (Gal et al., 2017; Pinsler et al., 2019) due to their probabilistic output. However, in regressions, for AL, statistics have to be derived like in Buyu and Vittorio (2017); Kendall and Gal (2017) or a separate data analysis through task-invariant methods has to be developed (Sener and Savarese, 2018; Sinha et al., 2019; Yoo and Kweon, 2019). The benefit of the Bayesian approximation is that the epistemic and aleatoric uncertainties can be used to filter out the unlabelled pool of data. Apart from the uncertainty variances, the geometric acquisition function *CoreSet* (Sener and Savarese, 2018) is revised under the Bayesian methodology.

In principle, CoreSet treats the minimisation of the geometric bound between the objective functions of the annotated set $\mathbf{s}^0$ and of a repres-

entative subset $\mathbf{s}$ from the unlabelled examples. The bounding between the two losses applied in this context can be expressed as:

$$\left|\mathcal{L}_B(\mathbf{x}_i, \mathbf{y}_i \in \mathbf{s}) - \mathcal{L}_B(\mathbf{x}_j, \mathbf{y}_j \in \mathbf{s}^0)\right| \leqslant \mathcal{O}(\gamma_{\mathbf{s}}) + \mathcal{O}\left(\sqrt{\frac{1}{B}}\right), \tag{3.6}$$

where $\gamma_{\mathbf{s}}$ is the fixed cover radius over the entire data space and B is the budget of samples to annotate. It has been shown in Sener and Savarese (2018) that this risk minimisation can be approximated through the k-Centre Greedy optimisation problem Wolf (2011). As it relies on Δ, the $l2$ distances between samples of posterior distribution define the selection principle adapted from CoreSet under the following scope:

$$\arg\max_{i \in \mathbf{s}} \min_{j \in \mathbf{s}^0} \Delta(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j), \tag{3.7}$$

where $\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j$ are predicted skeletons corresponding from the initial subset $\mathbf{s}^0$ and the available unlabelled pool $\mathbf{s}$.

Considering the Bayesian DeepPrior architecture, the CoreSet solution is extended by including the epistemic variance in the distance computation from equation 3.7. Moreover, the estimated 3D coordinates are averaged after M MC Dropout inferences. Therefore, when evaluating $\min_{j \in \mathbf{s}^0} \Delta(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j)$, their corresponding standard deviations $\hat{\sigma}_{ep_i}$ and $\hat{\sigma}_{ep_j}$ are subtracted so that closer uncertain centres are highlighted. Implicitly, $\hat{\sigma}_{ep_i}$ and $\hat{\sigma}_{ep_j}$ extend these minimum pairwise distances ($\min_{j \in \mathbf{s}^0} \Delta(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j)$) the maximum distance from the unlabelled is computed. Finally, to the subset $\mathbf{s}^0$ the furthest uncertain samples are added for annotation. The impact of the epistemic uncertainty variance is adjusted with a parameter η. This number of steps is repeated according to a budget.

This adapted combination between the k-Centre Greedy algorithm and the epistemic variances as **CKE**. The pseudo-code from Algorithm 1 presents the steps for selecting a subset $B_{\mathbf{ublb}}$, given a budget B. Furthermore, in Figure 3.2 the data selection stages are conceptually represented for both CoreSet and CKE. It can be denoted that CKE benefits of the Bayesian model uncertainty $\hat{\sigma}^2_{ep}$ when minimising the global geometric cover $\gamma_{\mathbf{s}}$. In this manner, the proposed solution identifies hand poses furthest from the labelled centres and with the highest epistemic uncertainty deviation. Moreover, the new objective function with learnt aleatoric uncertainties (see Equation 3.2) drops the noisy samples making the estimations more robust. On these premises, it is considered that the AL method for Bayesian DeepPrior is superior to the standard approach.

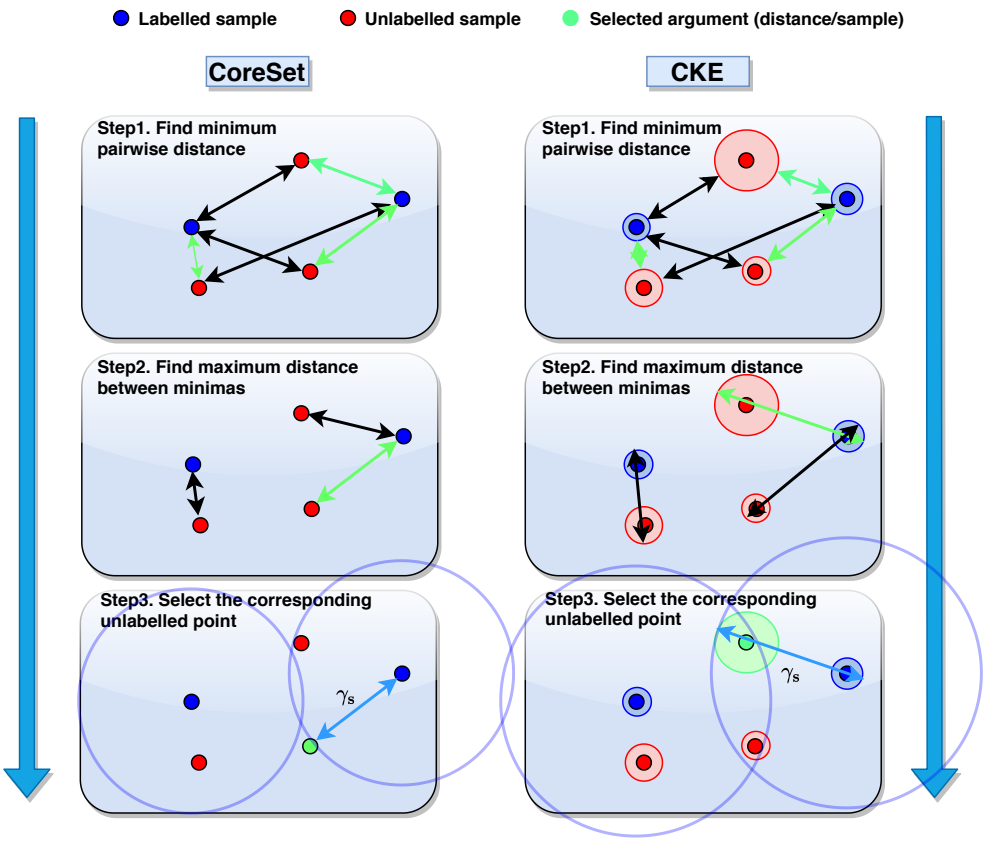

Figure 3.2: Conceptual comparison between the sampling techniques: Corset (Sener and Savarese, 2018) vs CKE. The blue dots represent the annotated samples in the hand skeleton space, while the red ones are to be selected. Each sample's uniform distribution defines the circle radii in the CKE process stages.

Algorithm 1 CKE

Input: labelled set $\mathbf{x}_j \in \mathbf{s}^0$, unlabeled pool $\mathbf{x}_i \in \mathcal{U}_{pool}$, query budget B, corresponding epistemic variances $\hat{\sigma}_{ep_i}$ and $\hat{\sigma}_{ep_j}$
Initialise $\mathbf{s} \subset \mathcal{U}_{pool}$
repeat
  $\Delta\mathbf{d}_{\mathbf{ub}_i} = \min_{j \in \mathbf{s}^0} \Delta(\hat{\mathbf{y}}_i + \frac{\eta}{2}\hat{\sigma}_{ep_i}, \hat{\mathbf{y}}_j + \frac{\eta}{2}\hat{\sigma}_{ep_j})$
  $\mathbf{arg}_{\mathbf{lb}_i} = \arg\min_{j \in \mathbf{s}^0} \Delta(\hat{\mathbf{y}}_i - \frac{\eta}{2}\hat{\sigma}_{ep_i}, \hat{\mathbf{y}}_j - \frac{\eta}{2}\hat{\sigma}_{ep_j})$
  $b = \arg\max_{i \in \mathbf{s}} \Delta\mathbf{d}_{\mathbf{ub}_i}(\mathbf{arg}_{\mathbf{lb}_i})$
  $\mathbf{s}_{\mathbf{ublb}} = \mathbf{s}^0 \cup \{b\}$
until $\mathbf{s}_{\mathbf{ublb}} = B + \mathbf{s}^0$
Return: $B_{\mathbf{ublb}} = \mathbf{s}_{\mathbf{ublb}} \setminus \mathbf{s}^0$

3.4.1 *3D Hand Pose Estimation Datasets*

To perform the Bayesian DeepPrior architectural choice together with CKE, a reiteration of the depth-based datasets is presented. Although the annotation process can be challenging to deploy automatically under noisy environments, there has been a detailed cumulative expansion of the hand data acquisition process. Therefore, the experiment section approaches three well-known hand datasets: ICVL (Tang et al., 2014), NYU (Tompson et al., 2014) and BigHand2.2M (Yuan, Ye, Stenger, Jain and Kim, 2017).

ICVL. This is one of the earliest depth-based hand datasets. It consists of a total number of 17,604 with 16 joint annotations. Thus, it is divided between 16,008 training and 1596 testing images. While there are ten different subjects used in the acquisition process, ICVL (Tang et al., 2014) still lacks a high articulation degree under a single frontal viewpoint. However, the annotation process has been done with 16 hand joints tracking and manual tuning by keeping a relatively low labelling noise.

NYU. Created from 2 subjects, NYU has 72,757 training images and an 8,252 testing set. This benchmark has been created similarly to ICVL, but only two subjects were considered. It also presents various hand articulations and rotations at a 640x480 frame size. The hand skeleton is mapped on 36 joint locations. In this analysis, training and testing are done only on the frontal viewpoint, as in ICVL.

BigHand2.2M. It is the largest depth-based hand pose dataset, with 2.2 million of frames from 10 different hand models. The 21 key-point annotations have been automatically generated through the attached sensors. Compared to the two other datasets, BigHand2.2M (Yuan, Ye, Stenger, Jain and Kim, 2017) has a holistic viewpoint perspective with complex hand gestures. However, it shares the same image resolution as NYU. From a practical perspective, it is decided to uniformly sub-sample BigHand2.2M every ten frames. This filtering is applied due to numerous redundant frames from the continuous acquisition process. Moreover, in Yuan, Ye, Stenger, Jain and Kim (2017), it has been shown that with a sixteenth of the dataset, the average error increases only by 3 mm compared to training with the entire set. Therefore, in the experiments, 251,796 samples will be used for training and 39,099 for testing.

For all the datasets, the number of joints is standardised to 21 3D locations. A pre-trained UNet (Ronneberger et al., 2015) is deployed to detect hands and centre-cropped them to the dimension of 128×128.

Bayesian DeepPrior variants	A	B	C
NYU Testing MSE [mm]	**22.48**	22.94	22.65

Table 3.1: Ablation evaluation of Bayesian DeepPrior

3.4.2 *Bayesian DeepPrior Architecture*

Ablation Studies on Architecture. To approximate a BNN using MC Dropout as described in Section 3.3.1.2, three probabilistic variants are proposed by dropping out different combinations of layers. These three strategies are as follows: **A** - convolutional feature extractor; **B** - centrally, on the last convolution and first dense layer; **C** - throughout the entire DeepPrior and evaluated the performance on NYU data set. The comparison of the performance on these configurations is summarised in Table 3.1. Accordingly, when applying MC Dropout only on the feature extractor (**A**), it yields the best performance. This is because in **C**, Dropout brings too much regularisation, while **B** locks low-level features in the first convolutional layers. The observed trend is similar to one reported in the Bayesian SegNet (Kendall et al., 2015). Adaptive Moment Estimation (Adam) (Kingma and Lei Ba, 2015) optimiser is deployed with a learning rate of 10^{-3}, a batch size of 128 and a total number of $M = 70$ MC Dropouts. These parameters remain constant for all three setups. Throughout all of the upcoming experiments, the maintained Bayesian DeepPrior is the variant where Dropout is present after the convolutional layers.

After identifying the optimal configuration of Dropouts, the number of variational inferences M has been varied under the same hyper-parameters setting. It was observed that increasing the value of M does not impact the performance, however, it adds computational complexity. Lowering the value of M to 40 does not impact the performance. Hence, this value is kept for the rest of the experiments.

Bayesian DeepPrior vs Standard DeepPrior. To demonstrate the effectiveness of the Bayesian DeepPrior approach, a quantitative comparison is conducted against the standard DeepPrior baseline proposed in Oberweger et al. (2015). The average MSE is evaluated for both train and test sets of the three compared benchmarks. The performance comparison is summarised in Table 3.2.

The Bayesian DeepPrior brings a clear advantage over Standard by yielding the lower testing error on all three benchmarks. This shows how effectively it can generalise over all testing sets, while the standard DeepPrior over-fits the training sets. In addition to the performance improvement, the Bayesian

Hand Dataset		DeepPrior	Bayesian DeepPrior
ICVL	train	7.1633	7.5261
	test	10.6233	**10.0988**
NYU	train	4.7034	7.7566
	test	25.0754	**22.4838**
BigHand2.2M	train	12.7731	7.7566
	test	22.2353	**21.4649**

Table 3.2: Bayesian DeepPrior vs DeepPrior - averaged MSE [mm]

Hand Dataset		aleatoric var	epistemic var
ICVL	train	1.057	0.004
	test	1.016	0.0039
NYU	train	1.169	0.0142
	test	1.512	0.0066
BigHand2.2M	train	1.964	0.0132
	test	2.1594	0.0174

Table 3.3: Averaged epistemic and aleatoric variances at $\times 10^{-2}$ on ICVL, NYU and BigHand2.2M

DeepPrior also generates uncertainty metrics for the hand poses.

Epistemic and Aleatoric Uncertainties. For every coordinate of the hand skeleton, the BNN estimates their aleatoric and epistemic deviation. Table 3.3 enlists the average of both uncertainties for all three benchmarks. The values are in the range of 0 to 1. This is due to the normalised UVD coordinates of the hand joints.

As deducted in Kendall et al. Kendall and Gal (2017), higher aleatoric variances are consistently obtained than the epistemic ones. Hence, the noise present in the hand data tends to overcome the model's learning capability. Apart from BigHand2.2M, the testing set's epistemic variance seems lower than that of the training set. Furthermore, it can be noticed an increased aleatoric variance on BigHand2.2M compared to the other sets. A reason for it could be the noise during annotation.

3.4.3 *Active Learning Evaluation*

AL has shown to be an effective tool (Sener and Savarese, 2018,?; Sinha et al., 2019) in acquiring representative data for a learning model. Given the proposed Bayesian DeepPrior architecture, the CKE query method was presented

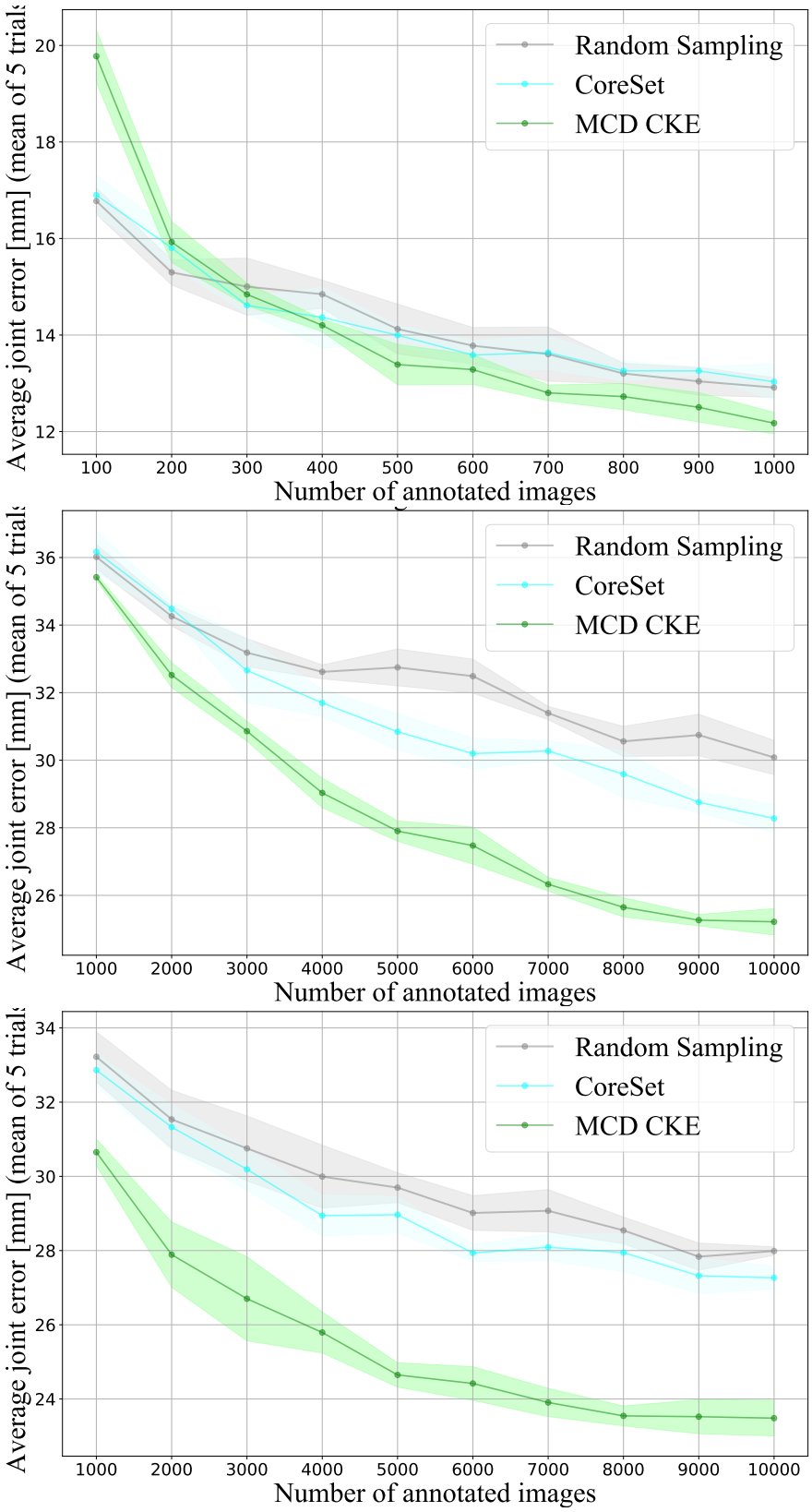

Figure 3.3: Empirical comparison I: Quantitative analysis of the proposed CKE method with Bayesian DeepPrior against the other methods applied on the standard version. Evaluated datasets: ICVL (Top), BigHand2.2M (Middle) and NYU (Bottom).

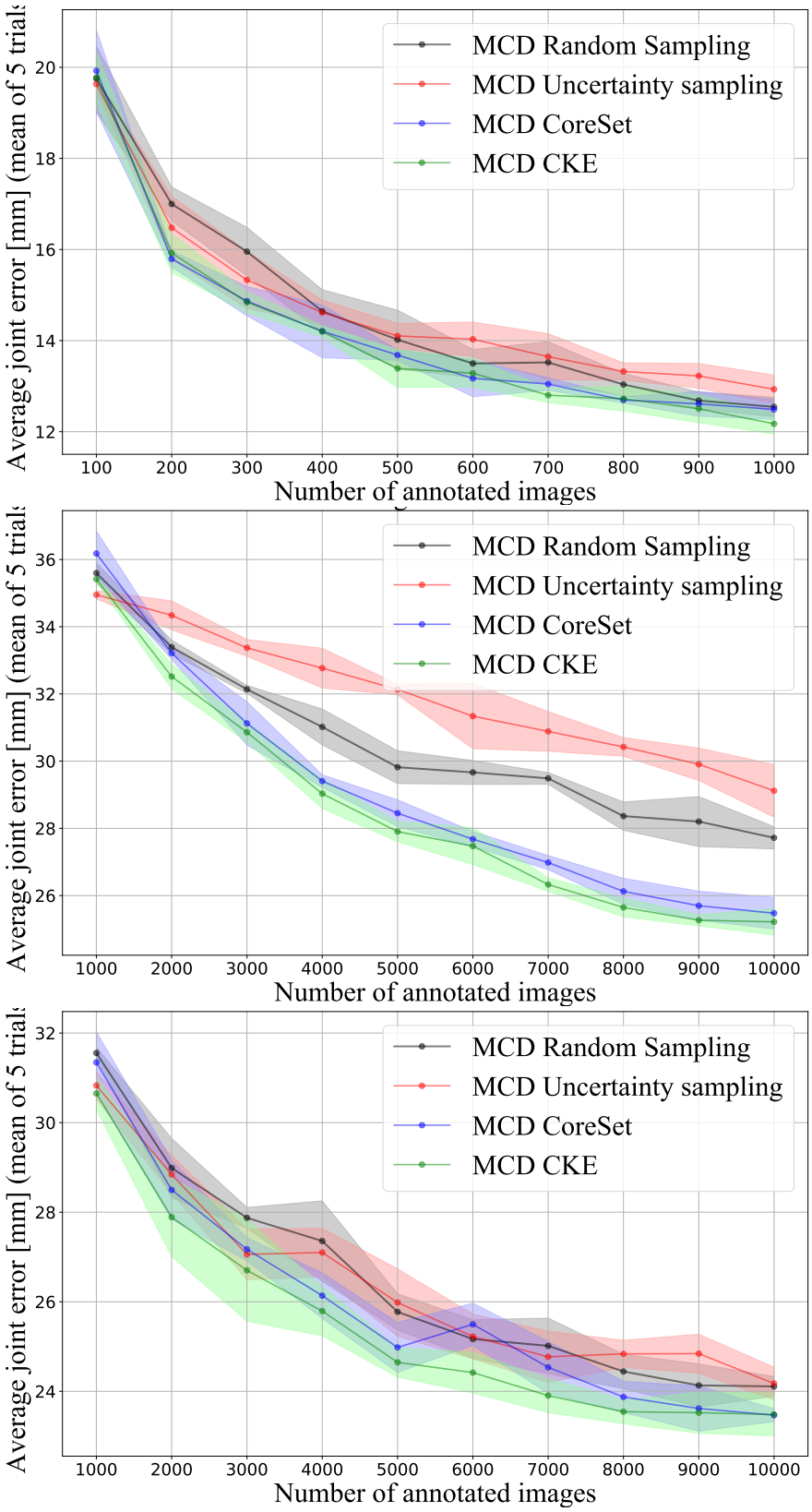

Figure 3.4: Empirical comparison II: AL performance comparison of the proposed CKE method against other MC Dropout adapted samplers on ICVL(Top), BigHand2.2M(Middle) and NYU(Bottom) data set.

in Section 3.3.2.2. While the pipeline gets an advantage from reducing the noisy samples through learnt aleatoric variances, CKE gathers both geometric and epistemic information regarding new unlabelled poses. AL selection baselines used for both Bayesian and standard DeepPrior architecture are presented accordingly:

Random sampling: The typical approach of annotating data just by uniformly sampling the unlabelled pool $\mathcal{U}_{pool}$.

Uncertainty sampling: Although for 3D-HPE regression, there is no confidence measurement like in classification tasks, this method is applied only to the evaluated epistemic uncertainties. The unlabelled examples are inferred through the Bayesian DeepPrior, and their epistemic variances result after 40 MC Dropouts. For each predicted skeleton, all the corresponding epistemic deviations are summed up and the topmost uncertain hand poses are selected for annotation.

CoreSet: It is one of the current state-of-the-art geometric acquisition function. Also, it has been widely used due to its task-agnostic properties relying either on the model's extracted features or on the output space. In this evaluation, CoreSet is directly applied to the predicted skeletons of both labelled and unlabelled sets.

CKE: This is the proposed method that extends the CoreSet functionality by including the epistemic deviations in the risk minimisation between the loss of the newly selected samples and the loss of the available labelled set. The approached baselines have been state-of-the-art in AL at the time of the proposed method. Other, more recent methods will be presented in the following chapters.

3.4.3.1 *Quantitative Results of Selecting Methods*

In this part, the three hand pose datasets are evaluated (ICVL, NYU and BigHand2.2M), under the pool-based scenario. A quantitative comparison is performed between the random sampling and CoreSet on the standard DeepPrior learner followed by CKE, the proposed selection method, on the Bayesian 3D-HPE. This follows the standard pool-based protocol, where there is a large pool, $\mathcal{U}_{pool}$ of unlabelled data. A small size initial set seed annotations $\mathbf{s}^0$ are made available for the first offline training of the target model. After this, the same practice is followed as in Beluch Bcai et al. (2018); Yoo and Kweon (2019) to randomly create a smaller pool $\mathbf{s} \subset \mathcal{U}_{pool}$. Thus, the selection methods are efficiently deployed on $\mathbf{s}$ under the budget B. This loop is systematically repeated over ten stages. As a quantitative metric, the performance is measured in averaged MSE. Also, due to the variation in the

initial selected set $\mathbf{s}^0$ and new random subsets $\mathbf{s}$, the results are averaged over five trials and also it computes the standard deviations.

For NYU and BigHand2.2M, the budget (B) is set similarly to the seed annotations $\mathbf{s}^0$ at 1000 samples. Whereas, for ICVL, due to its smaller size, it is set at 100. The intermediate subset $\mathbf{s}$ size used for AL selection is set to 10% of the entire U_{pool} which results in 20,000 of BigHand2.2M; 7,276 of NYU; and 1,601 of ICVL. In the CKE query function, the most informative samples were yielded when the uncertainty influence parameter was set to $\eta = 0.3$.

Figure 3.3 compares the performance of the proposed method with the default baseline (random sampling) and one of the state-of-the-art methods (CoreSet) in three different challenging data sets. Referring to the same figure, it can clearly be seen that the proposed method consistently outperforms the existing state-of-the-art. Specifically, after ten AL passes, the lowest averaged MSE is achieved on every benchmark accordingly: **12.17 mm** for ICVL, **23.48 mm** for NYU and **25.21 mm** for BigHand2.2M. This demonstrates how effective the Bayesian AL is in obtaining maximum accuracy with fractions of the entire datasets.

Similarly, Figure 3.4 illustrates the performance of the existing selection methods and the proposed sampling technique CKE when applied with the Bayesian DeepPrior. Comparing the performance of CoreSet when the learner is DeepPrior (Figure 3.3) vs when the learner is Bayesian DeepPrior (Figure 3.4), it can be noticed the improvement of the averaged MSE on every dataset. This trend is followed even in the random sampling technique. Hence, this demonstrates that modelling aleatoric and epistemic uncertainties while training learners in the AL framework is effective.

3.4.3.2 *Qualitative comparison*

The sampling methods are also evaluated qualitatively. The baselines (random and CoreSet) are compared against the proposed CKE function. To this end, the predicted 3D hand skeletons are extracted on the NYU test set and track their structural representation in the first two initial stages as well as in the last two AL stages. For comparison, three different poses are evaluated in the top, middle and bottom rows of Figure 3.5. In the left part, it can be observed that CKE, under the Bayesian DeepPrior learner, generates 3D hand poses quite closer to the ground truth from the early stages. In contrast, the other two methods failed to do so. These characteristics are equally visible even in highly articulated poses, as shown in the last row.

From a qualitative perspective, the relevance of the epistemic and aleatoric uncertainty values are also evaluated in the context of 3D joint error locations.

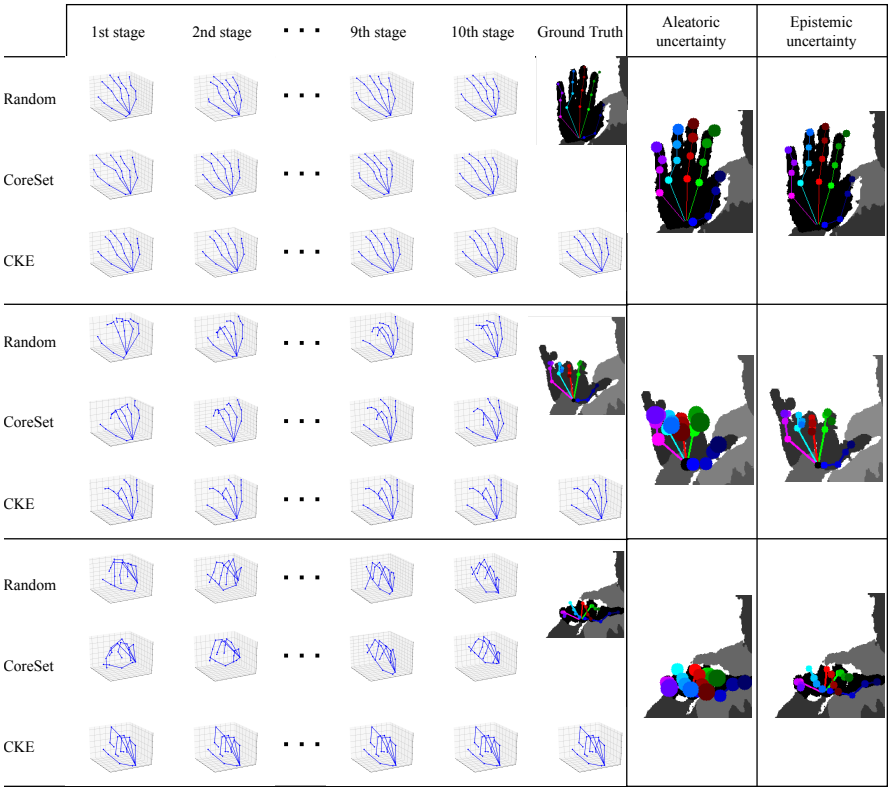

Figure 3.5: Qualitative comparisons on the Active Learning methods at different selection stages (Left). The 3D joint aleatoric and epistemic uncertainty variance (Right).

The last two columns of Figure 3.5 depict these uncertainties. The extended variance present at the fingertips (proportional to the radius of the circle) can be interpreted as high acquisition noise. Moreover, the occluded poses (middle and last row) have bigger circles showing that the model is less confident in such cases. This happens when there are not sufficient training examples in such extreme poses and occlusions. A more powerful estimator trained on annotated examples with such extreme poses and occlusion needs to be deployed to overcome this.

3.5 CONCLUSIONS AND LIMITATIONS TO ADDRESS

In this chapter, the first work of AL applied to the 3D-HPE task was presented. This method combines the most successful approaches at the time, uncertainty with MC Dropout and CoreSet, in an effective and state-of-the-art manner. Therefore, the pose estimator DeepPrior is approximated as a BNN while deriving its model and data-dependent uncertainties. It has been shown through qualitative and quantitative evaluation how the two uncertainties play a critical role in the AL scheme.

Furthermore, by combining the geometric sampling from CoreSet, the proposed sampling technique, CKE, is suitable for the Bayesian DeepPrior infrastructure. Under the pool-based scenario, the lowest 3D joint errors are achieved with the least amount of data for three well-known datasets. This chapter demonstrates the lack of representativeness and redundancy that can be present when gathering a 3D hand dataset. Therefore, a Bayesian approximation with the CKE acquisition method may help build a holistic and model-refined dataset while saving considerable annotation time.

While conducting the experiments with the three datasets, certain limitations of the methodology were encountered. One of them refers to the Bayesian approach and the way of extracting uncertainties. As mentioned above, the uncertainties are obtained after multiple MC Dropout inferences and averaging of a single input. This process adds an intense computational effort, which scales up with more complex architectures.

Another drawback consists of the intrusive architectural requirement of the Bayesian approximation. This requirement assumes that Dropout layers can be integrated easily in diverse DL models. However, it may not be necessary the case, and CKE limits its functionality. Other limitations concerning the performance of CKE are due to the 'cold-start' effect of the initial training set. CKE as CoreSet fully relies on describing the l2 space with the quality of the representations. Poor posteriors or latents of the learner will constrain the

diversity of selection. Some of those limitations are going to be addressed in the next chapters.

4

SEQUENTIAL GCN FOR ACTIVE LEARNING

4.1 COURSE OF OBJECTIVES WITH JUSTIFICATION

In this Chapter, some of the previous AL limitations are investigated by asking:

> "How can the acquisition function in DAL customise its selection in a more informative and task-aware manner?"

To shift the entire attention to AL, this Chapter proposes to ease the systematical sequential experiments of the pool-based scenario for other applications. The first and most convenient task is image classification. It has also been the first approached task in terms of DAL (Gal et al., 2017). The DL models also have fewer parameters, and this makes the experiments run faster.

By taking this advantage, however, the proposed AL methodology should be task-agnostic so that its selection can be transferred to the 3D-HPE task. The modular aspect of the AL method makes it independent of the learner's architecture. This approach encourages surpassing the limitations of CKE, which has been specifically designed on the predicted 3D hand skeletons.

The majority of the previous works (Abel and Louzoun, 2019; Beluch Bcai et al., 2018; Cai et al., 2017; Cortes and Vapnik, 1995; Gal et al., 2017; Gao et al., 2018; Wu et al., 2019) also integrate both the selection method and the learner together. Unlike these works, this Chapter proposes a parametric sampling methodology that is trained separately from the learner. This makes it task-agnostic, similarly to Sinha et al. (2019); Yoo and Kweon (2019).

GCN (Bronstein et al., 2017; Kipf and Welling, 2017) is a powerful tool to induce higher-order representations of nodes by performing message-passing operations between the neighbouring nodes. To answer the above question, the objective is to exploit the strength of GCN to discard redundant unlabelled samples for effective labelling. Thus, all the available data are represented in the form of a graph. Each node encodes an image description, and the edges give an idea of the similarities. In the beginning, a few examples are randomly annotated as before. These labelled examples act as seeds to propagate the labelled data information to the neighbouring nodes and identify unlabelled

look-a-like examples. Then, the parameters of the graph are learnt to minimise the binary-cross entropy between labelled and unlabelled examples.

As a first selection criterion, the examples are sorted based on the confidence scores through an uncertainty sampling approach. The labels of the most uncertain are queried and passed to both learner and sampler for another AL cycle. This enables the identification of unlabelled data manifolds which are redundant to the labelled examples. Recently, VAAL (Sinha et al., 2019) trained a VAE (Pu et al., 2016) in an adversarial manner to map images to a latent space where labelled vs unlabelled examples better discriminate. However, this method ignores the neighbouring relations between the examples and also there is no such mechanism for message passing.

As a second selection criterion, the node information from the GCN is gathered and distributed as in CoreSet (Sener and Savarese, 2018). An advantage over CoreSet is that the GCN representations add the propagated information between all the samples, and they also inherit uncertainty in the feature space.

The structure of this Chapter follows the previous one where the literature review is succeeded by the methodology and the list of experiments.

4.2 RELEVANT LITERATURE

4.2.1 *Model-based methods*

Recently, a new category of methods has been explored in the AL paradigm where a separate model from the learner is trained for selecting a sub-set of the most representative data. This method is based on this category. One of the first approaches (Yoo and Kweon, 2019) attached a loss-learning module so that the loss can be predicted offline for the unlabelled samples. In Sinha et al. (2019), another task-agnostic solution deploys a VAE to map the available data on a latent space. Thus, a discriminator is trained in an adversarial manner to classify labelled from unlabelled. The advantage of the proposed method over this approach is the exploitation of the relative relationship between the examples by sharing information through message-passing operations in GCN.

4.2.2 *GCN in Active Learning*

GCN (Kipf and Welling, 2017) has opened new AL methods that have been successfully applied in Abel and Louzoun (2019); Cai et al. (2017); Gao et al.

(2018); Wu et al. (2019). In comparison to these methods, this AL method has distinguished learner and sampler. It makes the approach task-agnostic and also gets benefits from the model-based methods mentioned just before. Moreover, none of these methods is trained sequentially. Wu et al. (2019) proposes K-Medoids clustering for the feature propagation between the labelled and unlabelled nodes. A regional uncertainty algorithm is presented in Abel and Louzoun (2019) by extending the PageRank Page et al. (1999) algorithm to the AL problem. Similarly, Cai et al. (2017) combines node uncertainty with graph centrality for selecting the new samples. A more complex algorithm is introduced in Gao et al. (2018) where a reinforcement scheme with multi-armed bandits decides between the three query measurements from Cai et al. (2017). However, these works (Cai et al., 2017; Wu et al., 2019) derive their selection mechanism on the assumption of a GCN learner. This does not make them directly comparable with this proposal where a GCN is trained separately for a different objective function than the learner.

4.2.3 *Uncertainty-based Active Learning methods*

As mentioned in Section 3.2, earlier techniques for sampling unlabelled data have been explored through uncertainty exploration of the CNNs (Gal et al., 2017; Houlsby et al., 2011; Kirsch et al., 2019; Pinsler et al., 2019). With the rise of GPU computation power, Beluch Bcai et al. (2018) ensemble-based method outperformed these variants. Despite this, querying in this manner is not advisable for some applications.

4.2.4 *Geometric-based Active Learning methods*

Although there have been studies exploring the data space through the representations of the learning model (Har-Peled and Kushal (2005); Kontorovich et al. (2016); Tsang et al. (2005)), the first work applying it for CNNs as an AL problem, CoreSet, has been presented in Sener and Savarese (2018). The key principle depends on minimising the difference between the loss of a labelled set and a small unlabelled subset through a geometric-defined bound. Furthermore, this baseline is represented again in the experiments as it successfully over-passed the uncertainty-based ones.

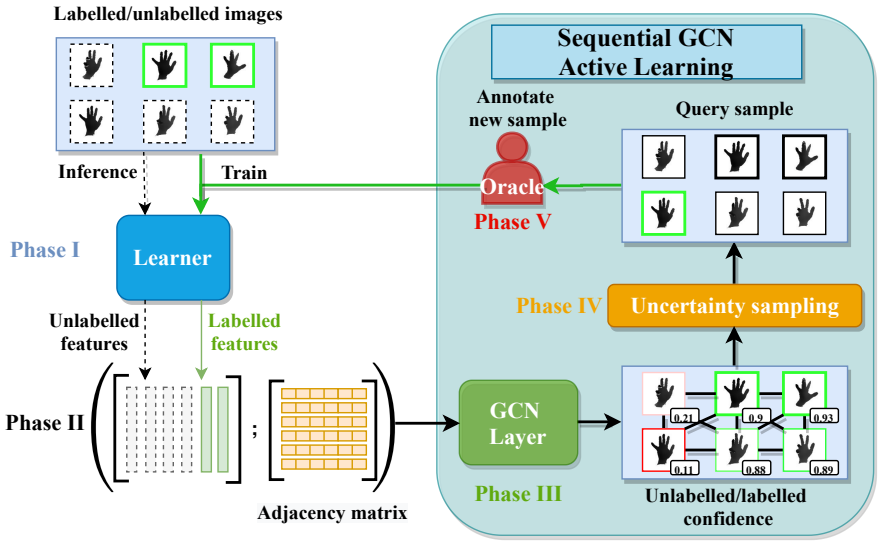

Figure 4.1: The diagram outlines the proposed pipeline. **Phase I**: Train the learner to minimise the objective of downstream task from the available annotations, **Phase II**: Construct a graph with the representations of images extracted from the learner as vertices and their similarities as edges, **Phase III**: Apply GCN layers to minimise binary cross-entropy between labelled and unlabelled, **Phase IV**: Apply uncertainty sampling to select unlabelled examples to query their labels, **Phase V**: Annotate the selection, populate the number of labelled examples and repeat the cycle.

4.3 METHOD

In this section, GCN-based sampler is described in detail. First, the learners are discussed in brief for the image classification and regression tasks under the pool-based active learning scenario. Two novel sampling methods are proposed in the second part: **UncertainGCN** and **CoreGCN**. UncertainGCN is based on the standard AL method uncertainty sampling Settles (2009) which tracks the confidence scores of the designed graph nodes. Furthermore, CoreGCN adapts the highly successful CoreSet Sener and Savarese (2018) on the induced graph embeddings by the sequentially trained GCN network.

4.3.1 *Learner*

In Figure 4.1, Phase I depicts the learner. Its goal is to minimise the objective of the downstream task. Both the classification and regression tasks are considered this time. Thus, the objective of this learner varies with the nature of the task.

Classification: A DL model $\mathcal{M}$ maps a set of inputs $\mathbf{x} \in \mathbf{X}$ to a discriminatory space of outputs $\mathbf{y} \in \mathbf{Y}$ with parameters θ. One of the renowned architectures,

ResNet-18 (He et al., 2016) is taken as the CNN model due to its relatively higher performance in comparison to other networks with comparable parameter complexity. Another model, VGG-11(Simonyan and Zisserman, 2015) has also been deployed in Appendix B. A loss function $\mathcal{L}(\mathbf{x}, \mathbf{y}; \theta)$ is minimised during the training process. The objective function of the classifier is cross-entropy, defined as below:

$$\mathcal{L}^{c}_{\mathcal{M}}(\mathbf{x}, \mathbf{y}; \theta) = -\frac{1}{N_l} \sum_{i=1}^{N_l} \mathbf{y}_i \log(f(\mathbf{x}_i, \mathbf{y}_i; \theta)), \tag{4.1}$$

where N_l is the number of labelled training examples and $f(\mathbf{x}_i, \mathbf{y}_i; \theta)$ is the posterior probability of the model $\mathcal{M}$.

Regression: To tackle the 3D-HPE, the same *DeepPrior* Oberweger et al. (2015) architecture from Chapter 3 is selected as model $\mathcal{M}$. Similarly, the 3D hand joint coordinates are regressed from the hand depth images. Thus, the objective function of the model changes as in Equation 4.2. In the Equation, J is the number of joints to construct the hand pose.

$$\mathcal{L}^{r}_{\mathcal{M}}(\mathbf{x}, \mathbf{y}; \theta) = \frac{1}{N_l} \sum_{i=1}^{N_l} \Big(\frac{1}{J} \sum_{j=1}^{J} \|\mathbf{y}_{i,j} - f(\mathbf{x}_{i,j}, \mathbf{y}_{i,j}; \theta)\|^2 \Big), \tag{4.2}$$

In order to adapt the method to any other type of task, the learner just needs to be modified. The remaining pipeline remains the same, detailed in the following sections.

4.3.2 *Sampler*

Moving to the second Phase from Figure 4.1, a pool-based scenario is adapted for AL. This follows the same principles presented in the previous Chapter, Section 3.3 and from the recent methods Beluch Bcai et al. (2018); Sener and Savarese (2018); Sinha et al. (2019); Yoo and Kweon (2019).

To review the scenario, from a pool of unlabeled dataset $\mathbf{D}_U$, an initial batch is uniformly sampled for labelling $\mathbf{D}_0 \subset \mathbf{D}_U$. Without loss of generality, in AL research, the major goal is to optimise a sampler's method for data acquisition, $\mathcal{A}$ in order to achieve minimum loss with the least number of batches $\mathbf{D}_n$. This scope can be simply defined for n number of AL stages as follows:

$$\min_{n} \min_{\mathcal{L}_{\mathcal{M}}} \mathcal{A}(\mathcal{L}_{\mathcal{M}}(\mathbf{x}, \mathbf{y}; \theta) | \mathbf{D}_0 \subset \cdots \subset \mathbf{D}_n \subset \mathbf{D}_U). \tag{4.3}$$

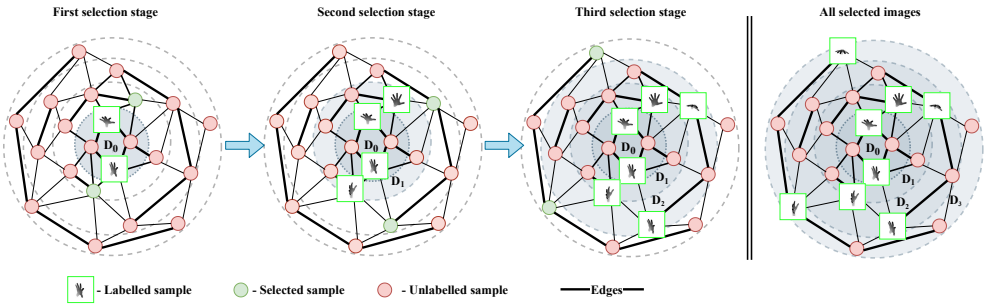

Figure 4.2: This Figure simulates the sampling behaviour of the **UncertainGCN** sampler at its first three selection stages. A toy experiment is run just by taking 100 examples from the ICVL (Tang et al., 2014) hand pose dataset. Each node is initialised by the features extracted from the learner, and the edges capture their relation. Each concentric circle represents a cluster of strongly connected nodes. Here, a group of images having similar viewpoints are in a concentric circle. Considering two annotated examples as seed examples in the centre-most circle, in the first selection stage, the algorithm selects samples from another concentric circle which is out-of-distribution than selecting the remaining examples from the innermost circle. Similarly, in the second stage, the sampler chooses images residing in another outer concentric circle which are sufficiently different from those selected before.

The aim is to minimise the number of stages so that fewer samples $(\mathbf{x}, \mathbf{y})$ would require annotation. For the sampling method $\mathcal{A}$, the heuristic relation between the discriminative understanding of the model and the unlabelled data space are brought together. This is quantified by a performance evaluation metric and traced at every querying stage.

4.3.2.1 *Sequential GCN selection process*

During sampling, as shown in Figure 4.1 from Phase II to IV, one of the contributions relies on sequentially training a GCN initialised with the features extracted from the learner for both labelled and unlabelled images at every AL stage. As stated before, similar to VAAL (Sinha et al., 2019), this methodology considers a model-based approach where a separate architecture is required for sampling. The motivation for introducing the graph is primarily in propagating the inherited uncertainty on the learner feature space between the samples (nodes). Thus, message-passing between the nodes after applying convolution on the graph induces higher-order representations of the nodes. Finally, the GCN will act as a binary classifier deciding which images are annotated.

Graph Convolutional Network. The key components of a graph, $\mathcal{G}$ are the nodes, also called vertices $\mathcal{V}$ and the edges $\mathcal{E}$. The edges capture the relationship between the nodes and are encoded in an adjacency matrix

A. The nodes $\mathbf{v} \in \mathbb{R}^{(m \times N)}$ of the graph encode image-specific information and are initialised with the features extracted from the learner. Here, N represents the total number of both labelled and unlabelled examples, while m represents the dimension of features for each node. After applying l_2 normalisation to the features, the initial elements of A result as vector product between every sample of $\mathbf{v}$ i.e. $(S_{ij} = \mathbf{v}_i^\top \mathbf{v}_j, \{i,j\} \in N)$. This indexes the similarities between nodes while falling under the same metric space as the learner's objective. Furthermore, from S, the identity matrix I is subtracted and then it is normalised by multiplying with its degree D. Finally, the self-connections are added back so that the closest correlation is with the node itself. This can simply be summarised:

$$A = D^{-1}(S - I) + I. \tag{4.4}$$

To avoid over-smoothing of the features in GCN (Kipf and Welling, 2017), a two-layer architecture is adopted. The first GCN layer can be described as a function $f^1_g(A, \mathcal{V}; \Theta_1) : \mathbb{R}^{N \times N} \times \mathbb{R}^{m \times N} \rightarrow \mathbb{R}^{h \times N}$ where h is the number of hidden units and Θ_1 are its parameters. A rectified linear unit activation Nair and Hinton (2010) is applied after the first layer to maximise feature contribution. However, to map the nodes as labelled or unlabelled, the final layer is activated through a sigmoid function. Thus, the output of f_g is a vector length of N with values between 0 and 1 (where 0 is considered unlabelled and 1 is for labelled). The entire network function can further be written as:

$$f_g = \sigma(\Theta_2(\mathrm{ReLU}(\Theta_1 A)A). \tag{4.5}$$

In order to satisfy this objective, the loss function will be defined as:

$$\begin{aligned} \mathcal{L}_g(\mathcal{V}, A; \Theta_1, \Theta_2) = -\frac{1}{N_l} \sum_{i=1}^{N_l} \log(f_g(\mathcal{V}, A; \Theta_1, \Theta_2)_i) - \\ - \frac{\lambda}{N - N_l} \sum_{i=N_l+1}^{N} \log(1 - f_g(\mathcal{V}, A; \Theta_1, \Theta_2)_i), \end{aligned} \tag{4.6}$$

where λ acts as a weighting factor between the labelled and unlabelled cross-entropy.

UncertainGCN: Uncertainty sampling on GCN. Once the training of the GCN is complete, the selection step can start. From the remaining unlabelled samples $\mathbf{D}_U$, their confidence scores $f_g(\mathbf{v}_i; \mathbf{D}_U)$ can be drawn as outputs of the GCN. Similarly to uncertainty sampling, **UncertainGCN** selects the

unlabelled images with confidence depending on a variable s_{margin}. While querying a fixed number of b points for a new subset $\mathbf{D}_L$, the following equation is applied:

$$\mathbf{D}_L = \mathbf{D}_L \cup \underset{i=1\cdots b}{\arg\max} |s_{margin} - f_\mathcal{G}(\mathbf{v}_i; \mathbf{D}_U)|. \tag{4.7}$$

This stage is repeated as long as Equation 4.3 is satisfied. In Section 4.4, an analysis for the choice of s_{margin} is provided, knowing the confident unlabelled scores should be closer to 0. Algorithm 2 summarises the GCN sequential training with the UncertainGCN sampling method.

Algorithm 2 UncertainGCN active learning algorithm

Given: Initial labelled set $\mathbf{D}_0$, unlabelled set $\mathbf{D}_U$ and query budget b
Initialise $(\mathbf{x}_L, \mathbf{y}_L), (\mathbf{x}_U)$ - labelled and unlabelled images
repeat
    $\theta \leftarrow f(\mathbf{x}_L, \mathbf{y}_L)$ ▷ Train learner with labelled
    $\mathcal{V} = [\mathbf{v}_L, \mathbf{v}_U] \leftarrow f(\mathbf{x}_L \cup \mathbf{x}_U; \theta)$ ▷ Extract features for labelled and unlabelled
    Compute adjacency matrix A *according to Equation 4.4*
    $\Theta \leftarrow f_\mathcal{G}(\mathcal{V}, A)$ ▷ Train the GCN
    for $i = 1 \rightarrow b$ **do**
        $\mathbf{D}_L = \mathbf{D}_L \cup \arg\max_i |s_{margin} - f_\mathcal{G}(\mathbf{v}; \mathbf{D}_U)|$ ▷ Add nodes depending on the label confidence
    end for
    Label $\mathbf{y}_U$ *given new* $\mathbf{D}_L$
until *Equation 4.3* is satisfied

CoreGCN: CoreSet sampling on GCN. To integrate geometric information between the labelled and unlabelled graph representations, a CoreSet technique (Sener and Savarese, 2018) is approached during the sampling stage. This has shown better performance in comparison to uncertainty-based methods (Wu et al., 2019). Sener and Savarese (2018) shows how bounding the difference between the loss of the unlabelled samples and the one of the labelled is similar to the *k-Centre* minimisation problem stated in Wolf (2011).

In this approach, the sampling is based on the l2 distances between the features extracted from the trained classifier. Instead of that, CoreGCN makes use of the GCN architecture by applying the CoreSet method to the features represented after the first layer of the graph. The sampling method is adapted to this mechanism for each b data point under the Equation:

$$\mathbf{D}_L = \mathbf{D}_L \cup \underset{i \in \mathbf{D}_U}{\arg\max} \min_{j \in \mathbf{D}_L} \delta(f^1_\mathcal{G}(A, \mathbf{v}_i; \Theta_1), f^1_\mathcal{G}(A, \mathbf{v}_j; \Theta_1)), \tag{4.8}$$

where δ is the Euclidean distance between the graph features of the labelled node $\mathbf{v}_i$ and the ones from the unlabelled node $\mathbf{v}_j$.

Finally, given the model-based mechanism, the proposed sampler is task-agnostic as long as the learner produces a form of feature representation. In the following section, we will experimentally demonstrate the performance of our methods quantitatively and qualitatively.

4.4 EXPERIMENTS

4.4.1 *Implementation details*

Hyper-parameters setting. For most of the experiments, ResNet-18 He et al. (2016) is set as learner due to its high accuracy and better training stability in comparison to the other contemporary architectures.

Moving to the training parameters, the batch size is set at 128. The models are optimised with SGD that has a weight decay 5×10^{-4} and momentum 0.9. At every selection stage, the model is trained for 200 epochs, starting with a learning rate of 0.1 and decreasing it to 0.01 after 160 epochs. These hyper-parameters are kept the same throughout all experiments. Similarly, the acquisition function (sampler) is approximated by the GCN trained sequentially. The chosen 2-layers GCN has a Dropout rate of 0.3 to avoid over smoothing (Zhao and Akoglu, 2020). The objective function is binary cross-entropy per node. The value of $\lambda = 1.2$ gives more importance to unlabelled points. For optimisation, Adam (Kingma and Lei Ba, 2015) is picked with a weight decay of 5×10^{-4} and a learning rate of 10^{-3}. The nodes of the graphs are initialised with the features of the images extracted from the learner network with the latest model parameters. And these features are projected to the dimension of 512. Such initialisation connects the learner with the acquisition function (sampler). Ablation studies were run to select these hyper-parameters which are presented shortly.

In the uncertainty-based sampling method, according to Equation 4.7, the s_{margin} variable has to be configured. From the ablation studies, it was observed that UncertainGCN selects the most informative samples when s_{margin} equals 0.1. This is intuitively correct as unlabelled samples with scores further from 0.1 will be selected.

Compared methods. The performance of the selection mechanism is evaluated against four comparable and state-of-the-art techniques:

Random: The most common practice is to uniformly sample b points for each subset of the entire unlabelled data set.

CoreSet(Sener and Savarese, 2018): The purpose of this method is to find b unlabelled samples that will act as new centres while the now reduced radius will cover the entire feature space from the learner.
VAALSinha et al. (2019): This work is considered one of the closest to the presented method. VAAL trains a VAE with both labelled and unlabelled samples to learn a discriminative subspace in an adversarial manner.

Learning Loss(Yoo and Kweon, 2019): It proposes to minimise an additional loss by learning the loss of the learner. The top samples with the highest loss will be sent for labelling
FeatProp(Wu et al., 2019): This GCN-based algorithm is chosen for comparison because it achieves the best performance on citation datasets against the recent AL frameworks for graphs (Abel and Louzoun, 2019; Cai et al., 2017; Gao et al., 2018). FeatProp first computes a distance matrix from the input features and then applies k-Means clustering. Consequently, it selects the unlabelled samples closer to the centroids. As this method expects pre-defined input features, in this scenario, FeatProp is still deployed as a sampling technique because it requires pre-defined input features. Furthermore, new features are generated with the ResNet-18 model at every selection stage.
Datasets and experimental settings. The proposed GCN method is tested on four challenging image classification benchmarks: CIFAR-10 (Krizhevsky, 2012), CIFAR-100 (Krizhevsky, 2012), FashionMNIST (Xiao et al., 2017) and SVHN (Goodfellow et al., 2013*b*). Each dataset has different properties and presents new challenges for the AL framework. CIFAR-10 consists of 50,000 images for training and 10,000 for testing. There are 5,000 samples for each of the 10 object categories. CIFAR-100 is constructed in a similar fashion with the exact size of the training and testing set. The difference is in the data distribution's granularity as 100 classes are categorised (500 images corresponding to each class). The SVHN dataset represents 10-digit classes in 73,257 train images and 26,032 test images.

Finally, FashionMNIST contains training and testing sets of 60,000 and 10,000, respectively, with annotations of 10 greyscale clothing designs.

For every dataset, all training images are assumed as unlabelled ($\mathbf{D}_U$) in the beginning. The initial labelled set $\mathbf{D}_L$ will be formed of 1,000 random sampled images for the CIFAR-10, SVHN and FashionMNIST experiments. Because of the fine-grained structure of CIFAR-100, during the first AL stage, 2,000 will be selected for labelling. These experiments are conducted along 10 stages, while at every stage the same number of queries will be produced by the proposed acquisition functions (1,000 b points for the 10-class datasets

and 2,000 for CIFAR-100). Similarly to the works in Beluch Bcai et al. (2018); Yoo and Kweon (2019), before selection, a randomly selected subset $\mathbf{D}_S \subset \mathbf{D}_U$ is created from the unlabelled images to avoid the redundancy occurrence common to those datasets. The $\mathbf{D}_S$ size will be 10,000 for all the AL stages in every experiment.

Evaluation Metrics. The mean average accuracy is computed for the 5 trials on test sets for every classification dataset.

4.4.2 *Quantitative comparisons*

Firstly, ResNet-18 is trained and tested using the entire training set for all the classification benchmarks. For CIFAR-10 and CIFAR-100, a classification rate of 93.09% and 73.02% is obtained, respectively. On the other two data sets, the learner yielded 93.74% on FashionMNIST and 95.35% on SVHN.

Figure 4.3 (top) shows the performance of UncertainGCN and CoreGCN against the other four existing arts on the CIFAR-10 dataset under the specified experiment settings. The solid line of the representation is the mean averaged accuracy, while the faded colour shows the standard deviation after the 5 trials. The GCN-based mechanism with the two sampling techniques surpasses random sampling and VAAL at every selection stage by at least 4%. After 10,000 labelled examples, the CoreGCN achieves state-of-the-art with 90.7%, the highest reported value in the literature at the time, (Sinha et al., 2019; Yoo and Kweon, 2019) under this configuration. The performance obtained with a fifth of $\mathbf{D}_U$ brings an important factor to the research community in whether labelling 40,000 more images is needed for a 2.4% gain. UncertainGCN comes close to CoreSet achievements over the testing set, but it is limiting in the latter stages due to the selection from confused feature-space areas. This aspect is demonstrated in the qualitative analysis.

Following the quantitative study on CIFAR-100 Figure 4.3 (bottom), it indicates that the proposed selections can scale to a more diverse dataset maintaining their state-of-the-art results. With 40% of labelled data, 69% accuracy is achieved by applying CoreGCN (4% less than from the entire dataset). Once again, the GCN capacity of passing information within the feature domain helps the active learning sampling furthermore. Compared to CIFAR-10, VAAL shows a better trend against random sampling, but it still falls under the presented performances. The reason is that the VAE might require a larger batch of queries (1,000 more) to differentiate. On the other side, although Learning Loss outperforms VAAL on the other datasets, for CIFAR-100, the mean averaged performance is similar.

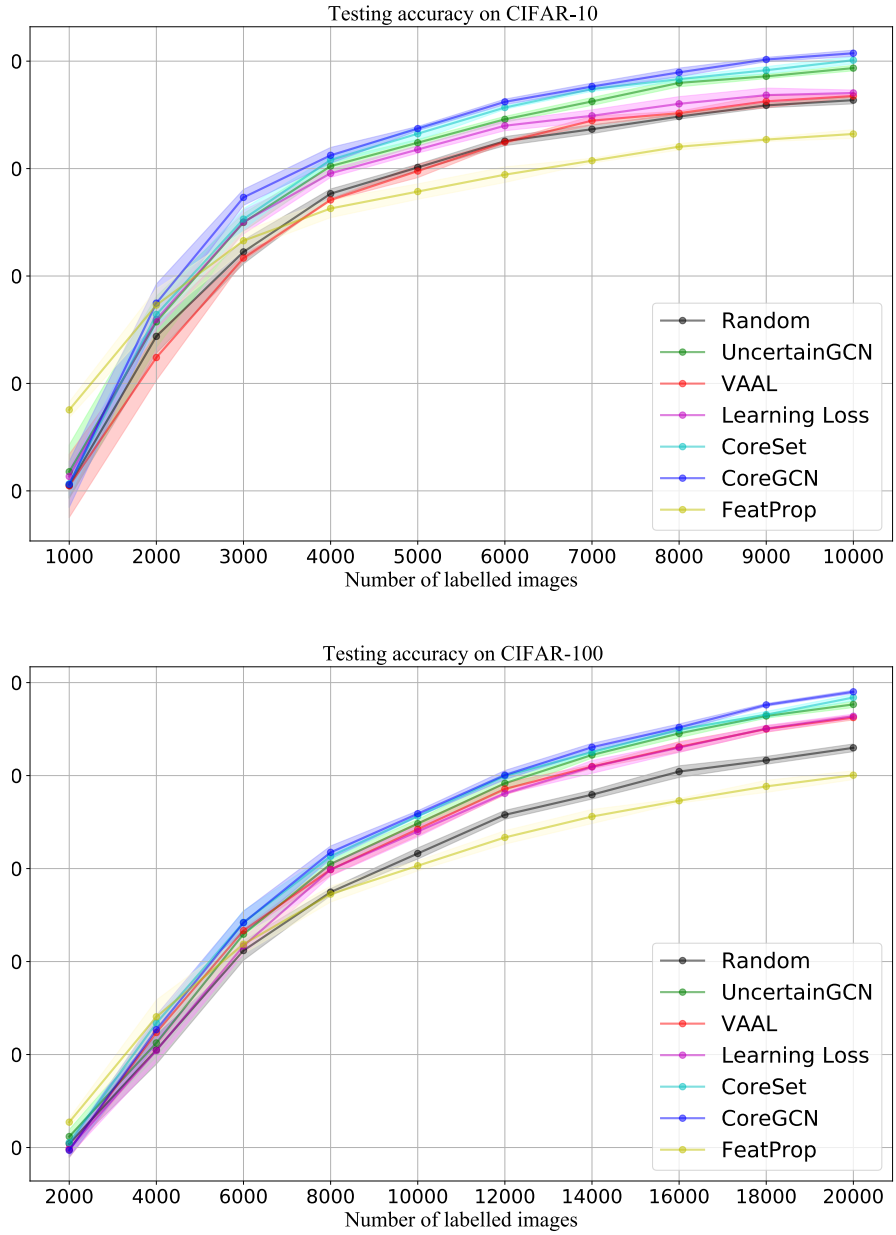

Figure 4.3: Quantitative comparison on CIFAR-10(Top) and CIFAR-100(Bottom)

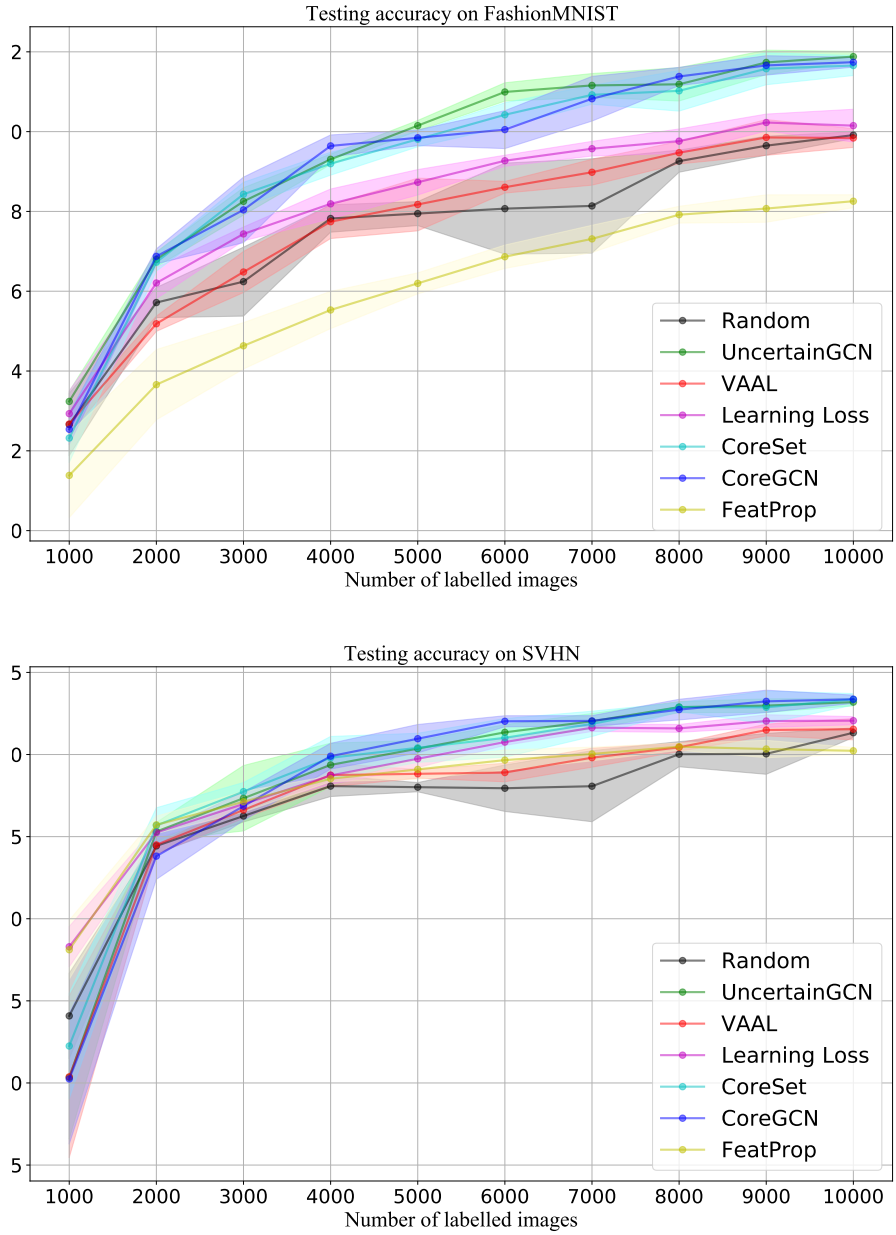

Figure 4.4: Quantitative comparison on FashionMNIST(Top) and SVHN(Bottom)

Continuing with the image classification analysis on FashionMNIST and SVHN, the performance progressions are illustrated in Figure 4.4. The challenge in those datasets consists in the high accuracy already obtained from the initial random set $\mathbf{D}_0$. This happens due to the ResNet-18's discriminatory capability and because of the saturation within the data space. In the FashionMNIST experiment, apart from VAAL, Learning Loss and FeatProp, all of the proposed methods follow similar curves over-passing random at every query stage. This shows that there are no apparent differences between the UncertainGCN and CoreGCN performances for greyscale data, especially with low-complexity features. The same behaviour is noticed for the SVHN dataset. The mean averaged accuracy plateaus after 6,000 labelled points around 92.5% for CoreGCN, while CoreSet and UncertainGCN follow up in the next stages. In this set of experiments, the GCN-based selection approach exceeds in terms of mean averaged accuracy the existing model-based methods like VAAL and Learning Loss.

From all datasets, it is clear that the FeatProp method, designed for GCN learners, is not suitable in image classification tasks. The accuracy shows promising results for the first selection stages, but then it gains minimal improvement on CIFAR-10, CIFAR-100 and FashionMNIST, falling behind random sampling. A similar plateauing effect can be noticed in Wu et al. (2019) on the PubMed dataset. Overall comparisons demonstrate the clear superior performance of the proposed method compared to the contemporary arts in DAL.

4.4.3 *Qualitative comparisons*

Ideally, the selection method aims to uniformly sample from all the classes at the initial annotation stages. Once the learner becomes more confident, the acquisition function should explore more samples from uncertain areas. Because in this AL framework, the exploitation factor is fixed to either 1,000 or 2,000, the exploration aspect is observed through t-distributed stochastic neighbor embedding (t-SNE) van der Maaten and Hinton (2008) representations.

The embeddings are computed from the ResNet-18 CIFAR-10 features of both labelled and unlabelled samples for the first and fourth query stage. In Figure 4.5, the t-SNE algorithms clusters the images of the 10 categories for both CoreSet and UncertainGCN. From an exploration perspective, UncertainGCN targets out-of-distribution samples and areas where features are hardly distinguished. The selected nodes have label confidence values

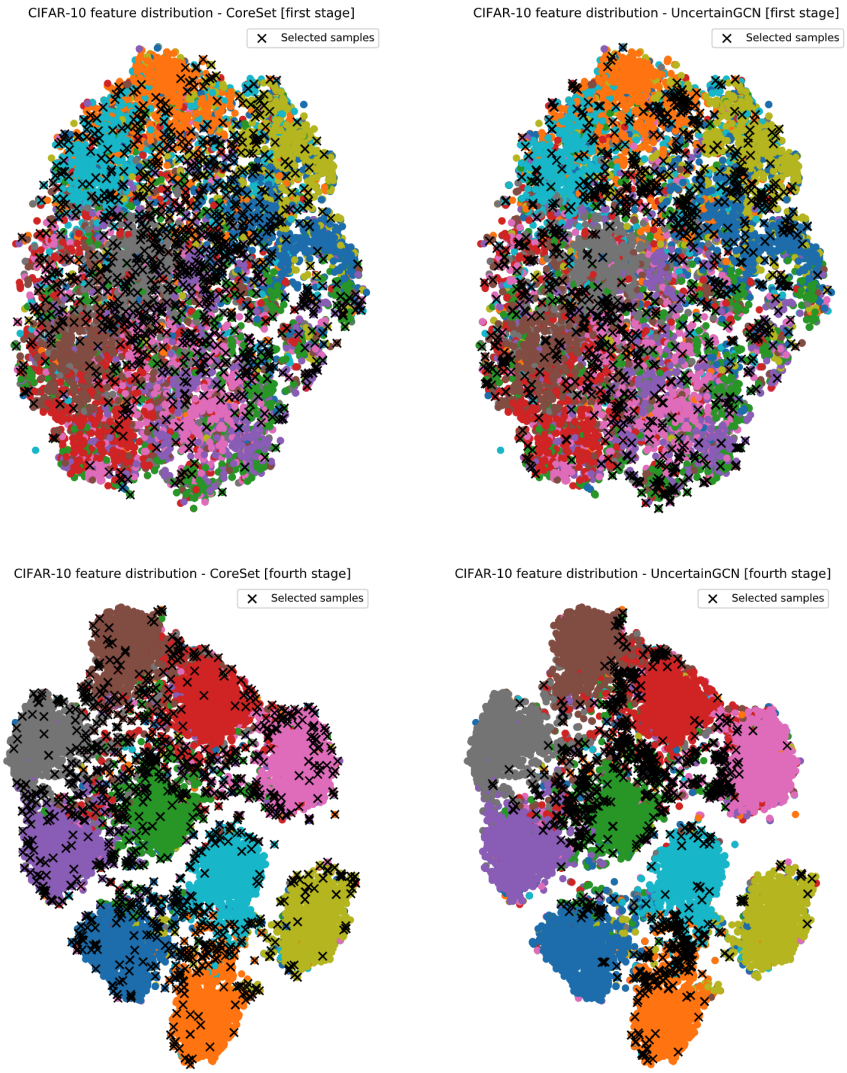

Figure 4.5: Exploration comparison on CIFAR-10 between CoreSet and UncertainGCN at the first (Top) and fourth (Bottom) AL selection stage

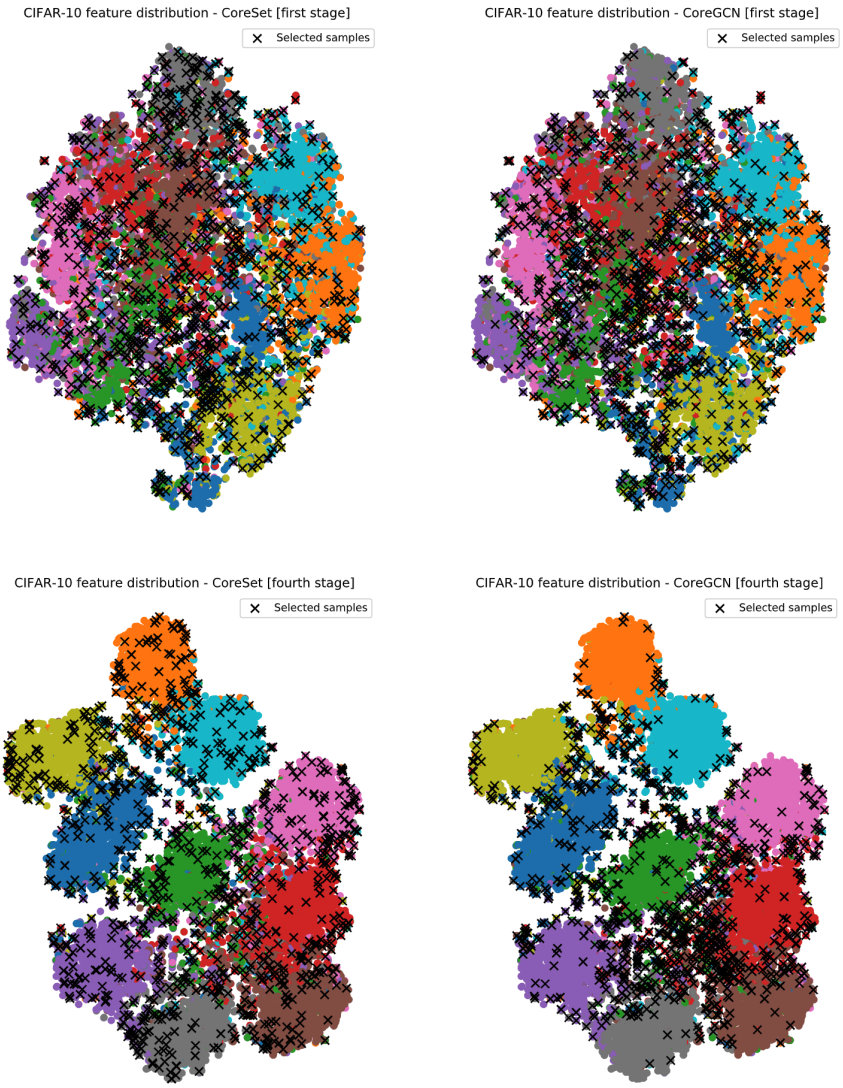

Figure 4.6: Exploration comparison on CIFAR-10 between CoreSet and CoreGCN at the first (Top) and fourth (Bottom) AL selection stage

higher than $s_{margin} = 0.1$. The reason is to select the least confident unlabelled samples that are found closer to 1. Furthermore, the uncertainty is inherited by the GCN binary classifier through its inability to correlate the multi-class nodes to the annotation information. As mentioned in the quantitative part, Figure 4.5 (bottom-right) shows the concentrated uncertain areas of this method. Since at first stages, the features are concentric, the GCN will sparsely spread the messages between the nodes.

In Figure 4.6, the qualitative investigation is continued for the CoreGCN acquisition method. It can be observed that starting at the first selection stage, CoreGCN avoids over-populated areas while tracking out-of-distribution unlabelled data. Compared to UncertainGCN, the geometric information from CoreGCN maintains a sparsity throughout all the acquisition stages. Consequently, it preserves the message passing through uncertain areas while CoreSet keeps sampling closer to cluster centres. This brings a stronger balance in comparison to CoreSet between in and out of distribution selection with the availability of more samples.

Comparison with CoreSet (Sener and Savarese, 2018). Concluding from the qualitative analysis, it can be stated that the features distributed through GCN training tend to get propagated in uncertain areas when new samples are selected with CoreGCN. However, this mechanism is not present in the CoreSet method while making it more variant to the feature representations. Therefore, the accuracy of CoreSet might decrease when changing the learner whilst CoreGCN will maintain its robustness with the GCN re-training (i.e. Figure 4.8).

4.4.4 *Ablation studies*

In this part, it is investigated the impact of the GCN hyper-parameters and its uncertainty-based selection variables on the performance. Furthermore, given the current AL framework, the effect of using different features can be studied by changing the learner's architecture from ResNet-18 to VGG-11 (Simonyan and Zisserman, 2015).

While varying the architectural parameters of the GCN binary classifier, a poorer selection was encountered with the increase of the Dropout rate from 0.3 to 0.5 or 0.8. However, when changing the size of the hidden units to 256 and 512, the UncertainGCN sampling was not affected on CIFAR-10. This might require further optimisation for different datasets, although robustness is being shown.

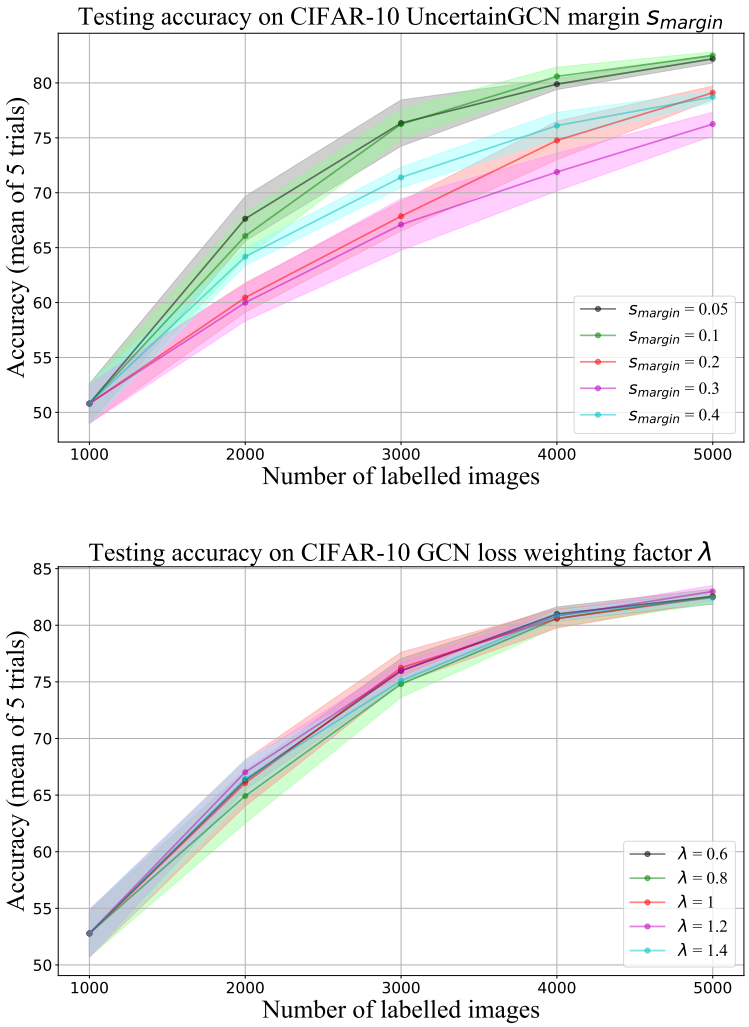

Figure 4.7: UncertainGCN tuning parameters

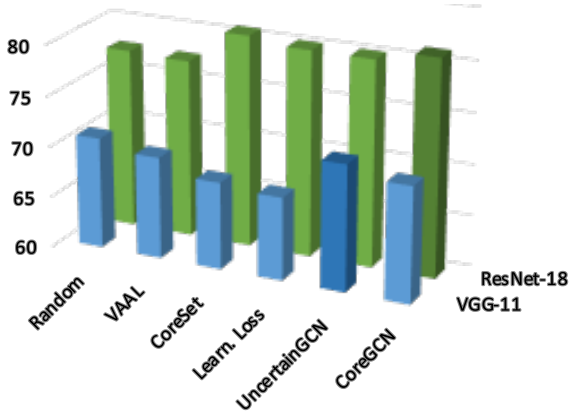

Figure 4.8: Learner comparison [Resnet-18 vs VGG-11] - 4,000 CIFAR-10 labelled images [mean averaged accuracy on 5 trials]

The main principle in UncertainGCN is focused on uncertainty sampling. In section 4.3.2.1, the parameter s_{margin} was defined to tune according to the GCN binary classification. Given the expected scores of 0 for confident unlabelled samples, the uncertain ones that yield scores closer to 1 are targeted. Figure 4.7 (top) confirms that by showing the best accuracy when s_{margin} is 0.1. Another key parameter in both UncertainGCN and CoreGCN query functions is λ, the loss bias between the labelled and unlabelled data points. In Figure 4.7 (right), the mean averaged accuracy on CIFAR-10 slightly improves in the first stages when the bias is shifted to the unlabelled loss by 20%.

In Figure 4.8, the architecture of the learner for CIFAR-10 experiment is changed to VGG-11 for analysing how the acquisition functions are affected in terms of accuracy up to the fourth sampling stage. In training the VGG-11 network, the same training hyper-parameters were kept as in ResNet-18. The proposed methods present robustness to this change, however, settings were left the same. Hence, these methods surpass all state-of-the-art at this early stage. This also demonstrates how the batch size and the feature representation play an important role in the performances of the other baselines. A representation of all selection stages is present in the Appendix B.

4.4.5 *Sub-sampling of Synthetic Data*

Unlike previous experiments of sub-sampling real images, the GCN-based method is applied to select synthetic examples obtained from StarGAN (Choi et al., 2018) trained on RaFD (Langner et al., 2010) face expressions manipulations. Although Generative Adversarial Networks (GANs) (Goodfellow et al.,

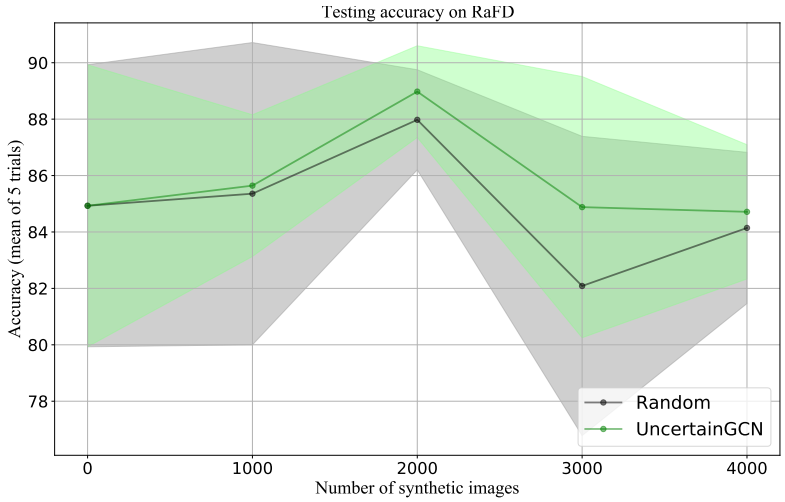

Figure 4.9: Performance comparison on sub-sampling synthetic data to augment real data for expression classification

2014) are closing the gap between real and synthetic data (Ravuri and Vinyals, 2019), the synthetic data and their associated labels are not yet suitable to train a downstream discriminative model. Recent study (Bhattarai et al., 2020) recommends sub-sampling the synthetic data before augmenting to the real data. Figure 4.9 shows the performance comparison of random selection vs UncertainGCN in 5 trials. From the experiments, the AL method outperforms the random selection by obtaining higher accuracy with less variance. The mean accuracy drops when the ratio of synthetic examples increases due to a large portion of GANs-synthetic examples being noisy (Bhattarai et al., 2020).

4.4.6 *3D Hand Pose Estimation*

The main task tackled in this thesis is revisited with the GCN sampler. This demonstrates once again the AL framework's adaptability to different computer vision tasks. Therefore, facing the same regression problem, only random sampling and CoreSet(Sener and Savarese, 2018) are chosen as baselines. Moreover, the latter has also shown greater performance in image classification than VAAL(Sinha et al., 2019), Learning Loss(Yoo and Kweon, 2019), and FeatProp(Wu et al., 2019). From an implementation standpoint, the fast-learning DeepPrior architecture defined in Oberweger et al. (2015) is also approached here. For all the experiments, the 3D-HPE is trained with the Adam (Kingma and Lei Ba, 2015) optimiser and a 10^{-3} learning rate over 128 batches. As in the previous Chapter, the images are pre-processed by centring and cropping with a hand detection algorithm based on U-Net

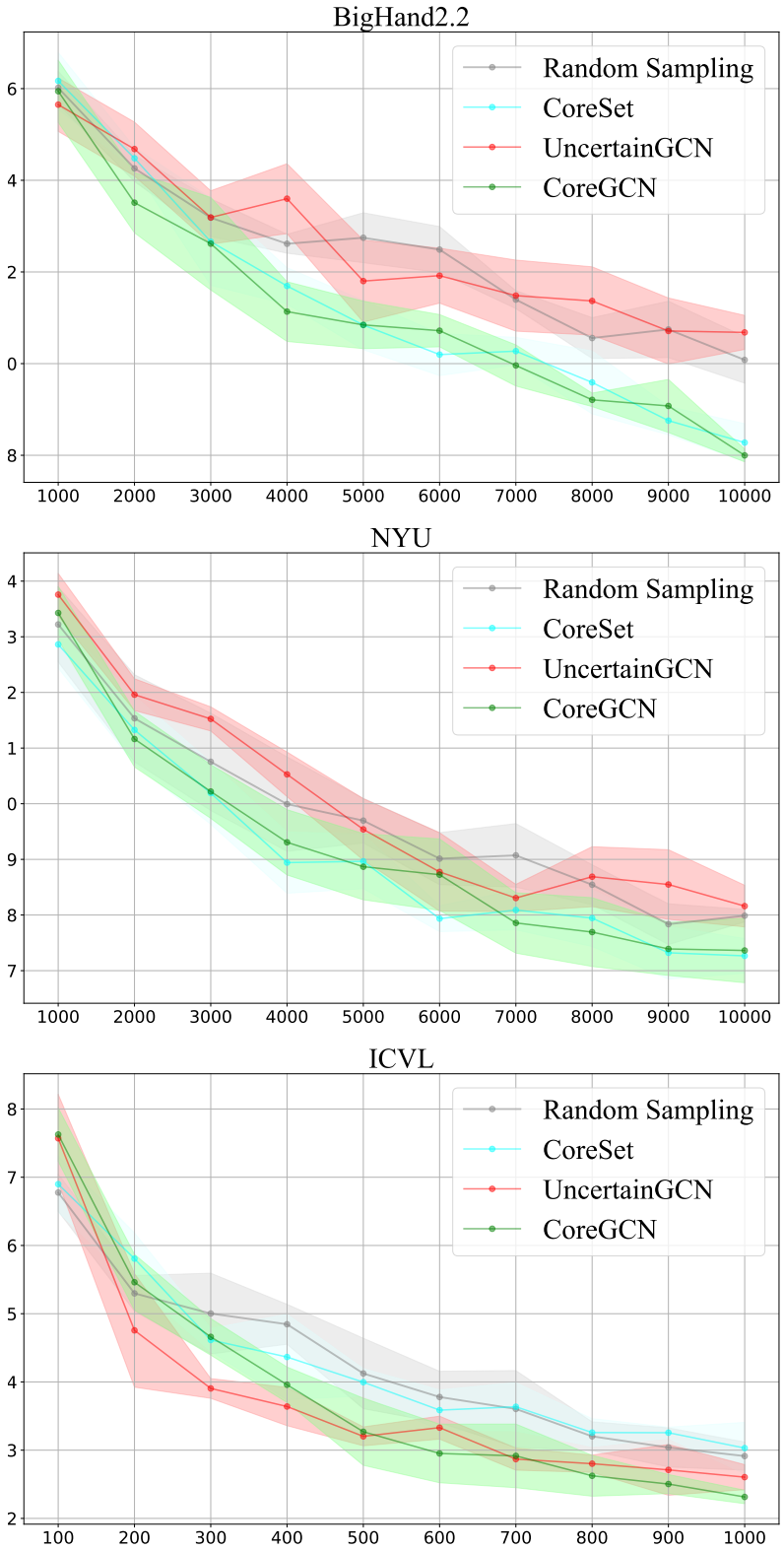

Figure 4.10: Quantitative comparison on 3D Hand Pose Estimation - BigHand2.2M(Top), NYU(Middle), ICVL(Bottom) (lower is better)

(Ronneberger et al., 2015). Furthermore, the baselines are evaluated on the same benchmarks: ICVL (Tang et al., 2014), NYU (Tompson et al., 2014), and BigHand2.2M(Yuan, Ye, Stenger, Jain and Kim, 2017). At every AL selection stage, the averaged testing set performance is measured through MSE in mm. All the other AL settings are kept as in Section 3.4.3.1.

After the ten selection stages, CoreGCN yields the lowest joint errors for all datasets, while UncertainGCN follows or outperforms CoreSet. Therefore, CoreGCN obtains a 12.3 mm error for ICVL, 27.2 mm for NYU and 28 mm for BigHand2.2M. Under the aforementioned settings, the main differences between the new methods and the other baselines can be observed on ICVL where UncertainGCN drops the error below 14 mm at only 300 annotated samples. Summing up, it is shown that the proposed AL framework can also effectively save the annotation time in the 3D-HPE.

4.5 LIMITATIONS AND CONCLUSIONS

In terms of computational efficiency, it can be acknowledged that the parameters of the GCN grow in the quadratic order to the number of nodes. This presents certain limitations which may be tackled either through prior sub-sampling of the unlabelled pool (as in this Chapter) or through inductive learning for sub-graphs (GraphSAGE (Hamilton et al., 2017), GraphSAINT (Zeng et al., 2020)).

As the GCN sampler is a parametric method, the number of the parameters was evaluated in comparison to VAAL and Learning Loss at the first query stage (as they fall within the same AL category). For selecting samples with VAAL, a VAE and a discriminator need to be trained in an adversarial manner. Thus, the total number of parameters for VAAL architecture is 23,115,204. Although it claims efficient sampling speed, the GCN sampling technique requires only 82,307. On the other hand, the Learning Loss module with 123,905 parameters has to be trained together with the learner affecting the overall training time. Moreover, specific DL architectures have to be deployed for this method where its complexity varies with the dimensions of the multi-features. Conclusively, the GCN-based sampler can be faster in certain scenarios when compared to the other model-based algorithms. All these AL frameworks scale up in terms of complexity with more labelled images.

Exclusively to the 3D-HPE and some image classification (SVHN and FashionMNIST) experiments, for all the AL baselines, the averaged performances do not have smooth curves with every new labelled set. This effect happens due to the distribution shift that the AL selection might cause. It is also highly

correlated with the 'cold-start' problem, which further shifts the objective after each exploration.

Despite these shortcomings to be overcome, this Chapter presented a novel parametric methodology for AL using GCN. Through comprehensive experiments, both quantitatively and qualitatively, the proposed selection variants, UncertainGCN and CoreGCN, produced highly competitive results for several tasks. This highlighted the task-agnostic behaviour where informativeness was maximised by GCN high-order representations. When extending to other tasks, it would be beneficial for developers to derive similar qualitative analyses 4.4.3 of the representations at different AL cycles. This would also help in deciding the configuration of the GCN and the type of the selection function (either uncertainty or geometric-based).

5

LEVERAGING UNLABELLED DATA FOR ACTIVE LEARNING SELECTION

5.1 COURSE OF OBJECTIVES WITH JUSTIFICATION

The third set of questions from the Introduction:

> "How can self-supervision help AL in dealing with the 'cold-start' problem and distribution shift ? Can the unlabelled examples be leveraged to improve the initial visual concepts of the learner?"

is planned to be answered in this chapter.

With the success of recent model-based AL samplers (Agarwal et al., 2020; Caramalau et al., 2021; Kim et al., 2021; Sinha et al., 2019; Yoo and Kweon, 2019), the increase of representativeness from both labelled and unlabelled samples has been shown to be a promising direction. Although the GCN-based sampler modelled the propagation of the feature representations, these representations are still bound to the learner's capability to discriminate at a specific stage. In an attempt to overcome this pool-based scenario constraint, SSL is approached in the context of AL.

SSL methods Berthelot et al. (2019); Caron et al. (2021); Grill et al. (2020); He et al. (2020); Xie et al. (2021) have made tremendous progress in generating discriminative representations of the images. Some of the methods have even come close to supervised methods in generalization (Chen, Fan, Girshick and He, 2020; Grill et al., 2020; Xie et al., 2021). One of the earliest AL works in this direction Gao et al. (2020) employed consistency loss between the input image and its geometrically augmented versions along with the objective of downstream tasks. However, this method limits augmentation methods in the primitive form. Similarly, Bengar et al. (2021) introduced contrastive learning in AL, but the self-supervised method and end-task objective are optimised in multi-stage form. This makes the model sub-optimal, affecting the features' representativeness during selection. Simple random labelling overpasses any AL criteria for this method. Thus, the existing works in this direction still show explicit limitations.

To address all the above issues, a novel component, *MoBYv2*, is introduced in the AL pipeline of this chapter. This component optimises the loss of

both the downstream task and a self-supervised loss jointly, as shown in Figure 5.1. The self-supervised loss is inspired by MoBY (Xie et al., 2021), which inherits the strength of two popular SSL methods: MoCo (He et al., 2020) and BYOL (Grill et al., 2020). MoBY is chosen because it addresses other previous methods' computational complexities and shortcomings such as SimCLR (Chen, Kornblith, Norouzi and Hinton, 2020) or BYOL (Grill et al., 2020). *MoBYv2AL*, the proposed SSL AL framework, has two branches (as shown in Figure 5.1), one updates with gradient (query encoder) and another with momentum (key encoder). The parameters of the momentum encoder are updated in slow-moving average with query one. Moreover, the memory bank of keys from the momentum encoder keeps long dependencies with several mini-batches. Apart from minimising a contrastive loss, another advantage consists in the asymmetric structure of BYOL that captures distances from mean representation. MoBYv2 culminates the AL selection with the concept-aware function, CoreSet. For clarification, MoBYv2 is the self-supervision in the semi-supervised pipeline of the learner, and MoBYv2AL consists of the entire AL framework with CoreSet selection.

5.2 RELEVANT LITERATURE

5.2.1 *Self-supervised Learning*

For the past years, a new pillar, SSL, has arisen in unsupervised environments with linked goals to AL. Learning generalised concepts from large-scale data is critical for further expansion to various vision applications. SSL can be divided in two approaches: consistency-based (Berthelot et al., 2019; Caron et al., 2021; Sohn et al., 2020; Tarvainen and Valpola, 2017) and contrastive energy-based (Chen, Kornblith, Norouzi and Hinton, 2020; Chen, Fan, Girshick and He, 2020; Grill et al., 2020; He et al., 2020; Li et al., 2021; Xie et al., 2021).

Consistency regularisation looks to preserve the class of unlabelled data even after a series of augmentations. For example, both MixMatch (Berthelot et al., 2019) and DINO (Caron et al., 2021) sharpen the averaged pseudo-labelled predictions. Conversely, contrastive learning generally demands pairs of positive and negative examples while optimising the similarity/contrast between them. Dual networks are usually deployed to evaluate these losses either within the batch (as in SimCLR (Chen, Kornblith, Norouzi and Hinton, 2020)) or within a dictionary of keys (for methods like MoCo (He et al., 2020), MoBY (Xie et al., 2021)). Because contrastive learning is foundational to this

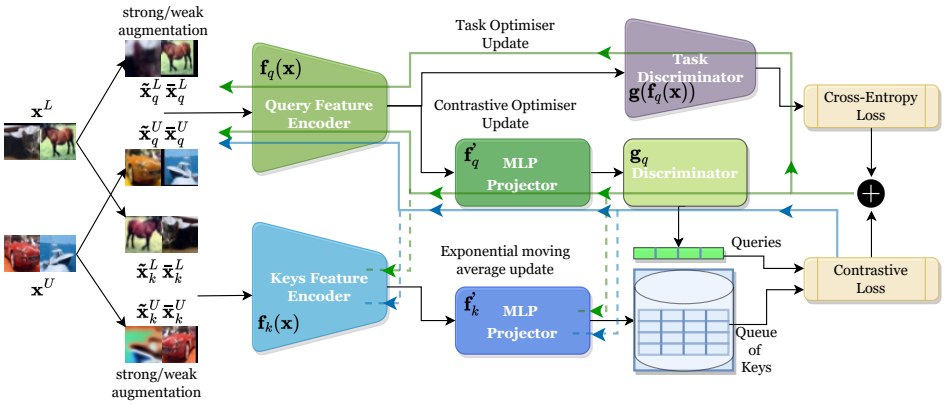

Figure 5.1: SSL-AL training framework under the proposed MoBYv2 configuration. The query feature encoder plays two roles: to map the features to the task discriminator for classification; to capture contrastive visual representation with the asymmetry of the query and key modules. For unlabelled data, the blue lines show the back-propagation of contrastive loss and its exponential moving average (dashed). The green lines also include the cross-entropy loss during training when the annotation is available.

proposal, these techniques are revised in Section 5.3.

5.2.2 *AL with self-supervision*

In the beginning, SSL and AL evolved in parallel for DL. Only recently, these fields have merged to further progress data sampling. Although SSL brings better visual constructs, there is still the question of which labelled information to allocate. By leveraging the unlabelled data behaviour, CSAL (Gao et al., 2020) firstly integrated MixMatch in the AL training and selection. The method proposed here follows a similar strategy, but it learns with end-to-end training contrastive representations. Despite this, CSAL is included in the SSL-based experiment as it is directly comparable.

Two new works tackle contrastive learning either in acquiring language samples, CAL (Margatina et al., 2021), or by adapting the sequential SSL SimSiam (Chen and He, 2021) in Bengar et al. (2021). CAL is task-dependent on NLP, and in Bengar et al. (2021), the multi-stage AL selection has no effect against random sampling. To this extent, these works are omitted in the experimental analysis.

The methodology approaches DAL as in the previous chapters. For simplicity, the primary task for the learner is image classification as in Chapter 4. As mentioned before, standard AL requires an online environment where the task learner selects and optimises simultaneously. A large unlabelled pool of data $\mathbf{D}_U$ is considered again, from which the initial subset is uniformly selected and labelled $\mathbf{S}_L^0 << \mathbf{D}_U$. Let $(\mathbf{x}^L, \mathbf{y}^L) \in \mathbf{S}_L^0$ be the available images and their corresponding classes. The learner is again a DL model comprising of a feature encoder $\mathbf{f}$ and a class discriminator $\mathbf{g}$. The objective loss for the learner is the categorical cross-entropy defined as $\mathcal{L}_{classification} = -\sum_{\mathbf{S}_L^0} \mathbf{y}^L \cdot \log \mathbf{g}(\mathbf{f}(\mathbf{x}^L))$.

Following the AL objective, the *exploration-exploitation trade-off* is decided in conjunction with the classification performance. Thus, the exploitation rate is set through a budget b across $\mathbf{D}_U / \mathbf{S}_L^0$ guided by a *selection criteria*. Consequently, the new sampled subset $\mathbf{S}_L^1$ is labelled for re-training the learner. The exploration factor is expressed by the *number of stages* $\mathbf{S}_L^{0\cdots N}$ that repeats this loop according to the targeted performance. While the exploration cycles may be limited, in this proposal, the primary focus is on exploitation.

5.3.1 *Contrastive Semi-supervised Learning framework*

5.3.1.1 *Self-supervised Learning method MoBY*

This Section briefly re-introduces the key aspects of the contrastive learning techniques combined in MoBY (Xie et al., 2021). These are constituents to the *MoBYv2* SSL proposal.

The goal of self-supervised learning aligns with the AL problem, where there is plenty of unlabelled data and a costly annotation procedure. However, the former tends to learn generalised visual representation in aid of the objective task. For contrastive learning, the main approach to obtaining these representations is by analysing the similarity (dissimilarity) within the data space. From the most successful works (Caron et al., 2021; Chen, Kornblith, Norouzi and Hinton, 2020; Grill et al., 2020; He et al., 2020; Xie et al., 2021), the contrastive learning process can be broadly formed of these main parts: data augmentation with or without dual encoder, feature-vector projections, similarity approximation by a dedicated loss function.

The self-supervision framework is designed according to *MoBY* (Xie et al., 2021). This method combines two innovative prior works *MoCo* (He et al.,

2020) and *BYOL* (Grill et al., 2020) on visual transformers (Dosovitskiy et al., 2020; Wu et al., 2020).

MoCo (He et al., 2020) pioneers contrastive learning by addressing the similarity between an image and a specific dictionary of samples. To deploy the loss, positive examples are required through data augmentation of the input query together with the other negative keys from the dictionary. The self-supervision training pipeline consists of two feature encoders and two Multi-Layer Perceptron (MLP) projectors for mapping the query and the keys. Consequently, the keys are permuted in a large memory bank, while the positive examples are inferred through the online encoder. The gradient over the dictionary of keys needs a slower update. Thus, a gradual momentum update is implemented.

BYOL (Grill et al., 2020), on the other hand, has a different approach to contrastive self-supervision. It simplifies MoCo by relying only on positive examples. In this way, the memory bank can be discarded. The InfoNCE (Oord et al., 2018) loss is also replaced with a l2 loss, given the new setting. The contrastive learning strategy of BYOL is indirectly obtained through batch normalisation. To achieve this, further modifications are proposed. Thus, the architecture of the dual encoders is asymmetric in regard to MoCo, and BYOL adds a prediction module to the projector of the online encoder. Following only positive examples, the inputs to the two networks are strong-augmented versions of the same image. Finally, BYOL preserves the common mode from the data and inherits contrastive learning when passing a slow exponential moving average from the online to the momentum encoder. MoBYv2 intuitively explores the contrastive learning strategies from both MoCo and BYOL and aligns the self-supervision with MoBY (Xie et al., 2021). The combined pipeline is depicted in 5.1.

From a design perspective, the asymmetric dual encoders are adopted from BYOL, as shown in Figure 5.1. The top branch culminates with a discriminator $\mathbf{g}_q$ to match the outputs from the bottom. Despite this, both branches consist of the same feature extractor architecture followed by an MLP projector ($\mathbf{f}_q^{'}, \mathbf{f}_k^{'}$ for query and key, respectively). Distinctively from MoBY, the CNNs are brought as feature encoders. Moreover, for optimisation, the MLP projectors and the query discriminator are reduced to a single layer with batch normalisation and ReLU activation.

The asymmetric pipeline helps to mimic the contrastive learning principle of BYOL. However, to include the concepts from MoCo, the objective is minimised with an adapted version of InfoNCE loss. In this case, the memory bank for the queue of keys still needs to be kept. The contrastive loss is

defined as a sum of InfoNCE from two augmented versions of a query $\{q, q'\}$ and of an other key image $\{k, k'\}$:

$$\mathcal{L}_{contrastive} = -\log \frac{exp(q \cdot k^{'}/\tau)}{\sum_{i=0}^{m} exp(q \cdot k_i'/\tau)} - \log \frac{exp(q^{'} \cdot k/\tau)}{\sum_{i=0}^{m} exp(q^{'} \cdot k_i/\tau))}, \quad (5.1)$$

where m is the size of the memory bank and τ is the adjusting temperature Wu et al. (2018). During training, the online query encoder branch is updated by gradient while the key encoder takes the slow-moving average with momentum. With this combined design, the preservation of both MoCo and BYOL representation concepts is ensured. On the one hand, the asymmetric structure indirectly finds discrepancies from the average image with moving average and batch normalisation. On the other hand, the contrastive loss with the queue of different keys maintains the direct distinctiveness between the images.

The standard SSL techniques, MoCo, BYOL, and MoBY, demand the supervision stage where the pre-trained models are fine-tuned for the task objective. Such multi-stage pipelines seem ineffective in AL (Bengar et al., 2021). In the next Section, the framework of *MoBY* is extended to minimise both the self-supervised objective and downstream task objective jointly.

5.3.1.2 *Semi-supervised Joint Objective*

A final step to clarify before presenting the joint training procedure is data augmentation. MoBY derives the augmentation strategy from BYOL, where the inputs suffer strong transformations. In the MoBYv2 proposal, an alternation between strong and weak augmentation is chosen, similar to MoCov2 (Chen, Fan, Girshick and He, 2020). This change boosted the performance of its predecessor (He et al., 2020). This was also observed in experiments that using only strong augmentations can affect the optimisation of the task-aware branch. The weak augmentations comprise of horizontal flips and random crops. In addition, the strong transformations include colour jitter (on brightness, contrast, saturation, hue), Gaussian blur, grayscale conversion and pixel inversion (solarise). From Equation 5.1, $\{q, k\}$ can be referred to as the weak transformations of query and key, and $\{q^{'}, k^{'}\}$ their corresponding stronger versions.

With all these elements in place, the learner from the existing AL pipeline is substituted with the query encoder, and it is trained jointly with the entire MoBYv2 pipeline. Starting from the first cycle, the available labelled

samples $(\mathbf{x}^L, \mathbf{y}^L) \in \mathbf{S}_L^0$ and the remaining unlabelled $\mathbf{x}^U \in \mathbf{D}_U$ are considered as queries and keys. A strong augmentation is marked as $\{\tilde{\mathbf{x}}_q^L, \tilde{\mathbf{x}}_k^L\}$, while a weak is represented with $\{\bar{\mathbf{x}}_q^L, \bar{\mathbf{x}}_k^L\}$. When training, the labelled and unlabelled batches are alternated with each inference. Therefore, only the contrastive loss is back-propagated for the data without annotation according to Equation 5.1. In this context, given the pipeline from Figure 5.1 for this contrastive loss $\mathcal{L}^U_{contrastive}(q, q^{'}; k, k^{'})$, $\{q, k\}$ and $\{q^{'}, k^{'}\}$ can be obtained so:

$$\{q, q^{'}\} = \mathbf{g}_q\,(\mathbf{f}_q^{'}\,(\mathbf{f}_q(\,\{\bar{\mathbf{x}}_q^U, \tilde{\mathbf{x}}_q^U\}))), \tag{5.2}$$

$$\{k, k^{'}\} = \mathbf{f}_k^{'}\,(\mathbf{f}_k\,(\{\bar{\mathbf{x}}_k^U, \tilde{\mathbf{x}}_k^U\})). \tag{5.3}$$

Here, $\mathbf{f}_q, \mathbf{f}_k$ are the query and key feature encoders, followed by their projectors $\mathbf{f}_q^{'}, \mathbf{f}_k^{'}$ and the asymmetric query discriminator $\mathbf{g}_q$.

Similarly, $\mathcal{L}^L_{contrastive}$, the contrastive loss, can be computed for the labelled images. In addition, the categorical cross-entropy, $\mathcal{L}_{classification}$, is also minimised with the output from the task discriminator. Once computed, both the contrastive and the classification loss are back-propagated. Therefore, the combined loss, adjusted by a scaling factor λ_c, can be expressed as:

$$\mathcal{L}^L_{combined} = \mathcal{L}_{classification} + \lambda_c \mathcal{L}^L_{contrastive}. \tag{5.4}$$

While the contrastive loss is computed continuously regarding the classification loss, we decide to reduce its influence over the gradients with $\lambda_c = 0.5$. Finally, the exponential moving average and the queue of keys are updated on the bottom branch for both labelled and unlabelled samples.

Up to this point, the MoBYv2 SSL method has been presented in reference to the learner of the AL pipeline. The distinguishable parts from its predecessor MoBY (Xie et al., 2021) are as follows: CNNs encoders; less number of layers for projectors and query discriminator; smaller key dictionary; different sets of augmentations inspired from MoCov2. Moreover, the training and setting of MoBYv2 is proposed with semi-supervision under a joint objective, while MoBY is solely trained with contrastive loss.

5.3.2 *Unlabelled samples selection*

For selection, it is important to emphasise that this proposal minimises the self-supervised loss inspired by MoBY. With this, the end-task objective jointly enriches the visual representations of the data compared to the DAL

strategies presented in the previous chapters. Consequently, AL selection methods that rely on the learner's data distribution will perform better.

CoreSet (Sener and Savarese, 2018) has been proven to be effective in such a scenario, and it is supported in all the previous experiments. To this extent, for selection, this geometric method is combined with MoBYv2. Briefly, CoreSet aims to find a subset of data points where a constant radius bounds the loss difference with the entire data space. This technique is approximated with k-Centre Greedy algorithm (Wolf, 2011) in the euclidean space of the feature encoder outputs $\mathbf{f}_q(\mathbf{x})$.

5.4 EXPERIMENTS

5.4.1 *Experiment settings*

Baselines. Random has a uniform selection at every stage, and it does not require extra processing steps. MC Dropout (Gorriz et al., 2017) creates a Bayesian approximation of the learner, and it samples data in regard to the modelled uncertainty. DBAL (Gal and Ghahramani, 2016) extends the same strategy but tracks uncertainty with entropy. From more recent methods that deploy dedicated modules to sample in the AL strategy, the baselines approached are Learning Loss (Yoo and Kweon, 2019), VAAL (Sinha et al., 2019), CoreGCN (Caramalau et al., 2021) and CDAL (Agarwal et al., 2020). The module in Yoo and Kweon (2019) learns to approximate the cross-entropy loss. VAAL pre-trains a VAE to discriminate between labelled and unlabelled data. CoreGCN is the GCN sampler with CoreSet from the previous chapter. The AL method proposed here, MoBYv2AL, follows CDAL in tackling the visual representation quality. Distinctively, CDAL captures contextual diversity information from repetitive spatial classes. The standard CoreSet selection is also added, which also influenced the of CDAL.

Datasets. For the quantitative evaluation, the same classification benchmarks from Chapter 4 are tested: CIFAR-10, CIFAR-100 (Krizhevsky, 2012), SVHN (Goodfellow et al., 2013*a*) and FashionMNIST (Xiao et al., 2017).

Models. In 5.3.1, it was mentioned that different CNNs are used for feature encoders. To show that MoBYv2AL is robust to architectural changes, VGG-16 (Simonyan and Zisserman, 2015) opts in the CIFAR-10/100 quantitative experiments and ResNet-18 (He et al., 2016) for SVHN and FashionMNIST.

Training settings. Training runs for 200 epochs at every selection stage, and the batch size is kept at 128. The dictionary size for the keys m is set up as in MoBY at 4096. In experiments, it was noticed that the contrastive and

cross-entropy loss converge together after 200 epochs. The learning rate starts at 0.01, and it follows a schedule for the queue encoder and task discriminator that decreases ten times at 120 and 160 epochs. However, the momentum scheduler update is kept in the key bottom branch (gradual momentum increment from 0.99). In the contrastive loss, for both queues, the temperature parameter is fixed to 0.2.

AL settings. Under the exploration-exploitation trade-off, the budget is characterised as an exploiting factor while the exploration is captured in the number of selection cycles. The initial random-sampled labelled dataset varies between the CIFAR-10/100 and SVHN/FashionMNIST experiments. For CIFAR-10/100, 10% (5000) of the entire training set is considered as labelled and the rest as unlabelled data. The budget is limited to 5% (2500) samples for selection and this cycle is repeated 7 times. In the second set of experiments, the method is tested in a more restrictive environment with a starting set of 1000 labelled and similar fixed budget. Despite this, the exploration is expanded to 10 cycles reaching 10000 labelled data. As a performance measurement, in the AL framework, averages of 5 trials testing accuracy are evaluated. These training and AL settings for CIFAR10/100 have been used in other AL works like VAAL (Sinha et al., 2019) and CDAL (Agarwal et al., 2020). Therefore, in the next experiments, these are comparatively included together with the other state-of-the-art baselines.

5.4.2 *Quantitative experiments*

CIFAR10/100. To maintain a fair comparison, in Figure 5.2, the performance charts obtained by CDAL (Agarwal et al., 2020) and VAAL (Sinha et al., 2019) are reported instead of re-running them. All methods use VGG-16 for the feature encoder. MoBYv2AL has a considerable advantage with the proposed SSL framework in the CIFAR-10/100 experiments from the first selection stage. The performance starts with a considerable gap from the other baselines due to the semi-supervised training scheme presented in 5.3.1.2. This shows the inherited contribution of the unlabelled contrastive representations in the query encoder.

In both CIFAR-10 and CIFAR-100, 20% testing accuracy is gained over standard learning (62% and 28% on CIFAR-10/100). This justifies the importance of the joint training framework from MoBYv2.

The SSL pipeline's more refined visual representations direct helpful information to the CoreSet selection method. Thus, a gradual increase is present in Figure 5.2, where after 7 cycles, with 40% labelled data, MoBYv2AL

achieves 89.6% mean accuracy on CIFAR-10 and 63.1% on CIFAR-100. Another observation in the CIFAR-10 experiment is that the AL performance saturates faster than in CIFAR-100. This effect occurs due to a large initial labelled pool in relation to the complexity of the task. MoBYv2 exploits more contrastive information, and it limits the exploratory potential in the next stages.

SVHN/FashionMNIST. From Figure 5.3, it can be deducted as well, that MoBYv2AL balances the exploration-exploitation trade-off when the initial labelled set is relatively low to the number of classes. The dark dashed line displays the supervised baseline training on the entire labelled set. While on CIFAR-10/100 and FashionMNIST, MoBYv2AL reaches comparable performance, by the end of the cycles, on SVHN, it surpasses after the sixth one (95%). Here, the relevance of the strong/weak augmentations is emphasised by enriching the discrete data distribution. Furthermore, greyscale data (as in FashionMNIST) can also benefit from the proposed AL framework. In Figure 5.3,the same results are kept of the previous experiments of UncertainGCN and CoreGCN. Even under these settings, MoBYv2AL outperforms with a noticeable consistent margin: for SVHN and FashionMNIST a gap of at least 2% - 3%.

Table 5.1: Comparison with the SSL-AL method CSAL on CIFAR-100 with a WideResNet-28 learner

SSL-AL method vs percentages of labelled	10%	15%	20%	25%	30%
CSAL	58.1	63.76	67.13	69.28	70.08
MoBYv2	**67.66**	**68.24**	**68.49**	68.57	**70.11**

Comparison with other SSL-AL. MoBYv2AL leverages unlabelled data for contrastive learning in the AL framework. In the previous experiments, this amount of data was equal to the available labelled samples. Therefore, at every AL cycle, this size increases with the newly selected data. Another recent SSL-AL baseline CSAL (Gao et al., 2020), however, deployed the consistency measurements from MixMatch (Berthelot et al., 2019) on the entire unlabelled data. It can be identified that MoBYv2AL over-exploits as CSAL the captured representation under these conditions. These two methods are compared on CIFAR-100 in Table 5.1 while the feature encoder is adjusted to WideResNet-28 (Zagoruyko and Komodakis, 2016). In this experiment, MoBYv2AL maintains the initial performance gain, but it is affected to some extent by saturation.

Imbalanced dataset experiment. Apart from SVHN, all the previous experiments have a uniform distribution over the classes. This rarely occurs

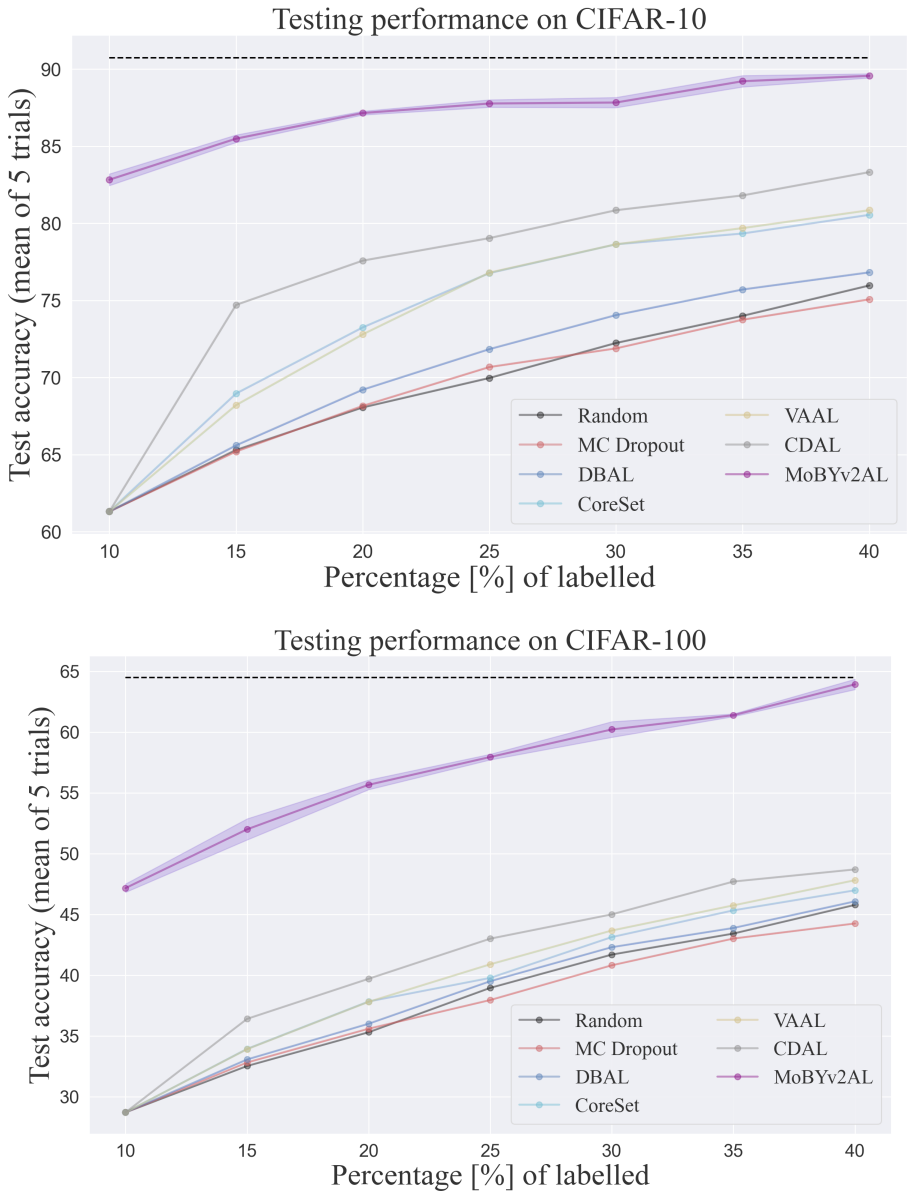

Figure 5.2: Evaluations on CIFAR-10 (Top), CIFAR-100 (Bottom)

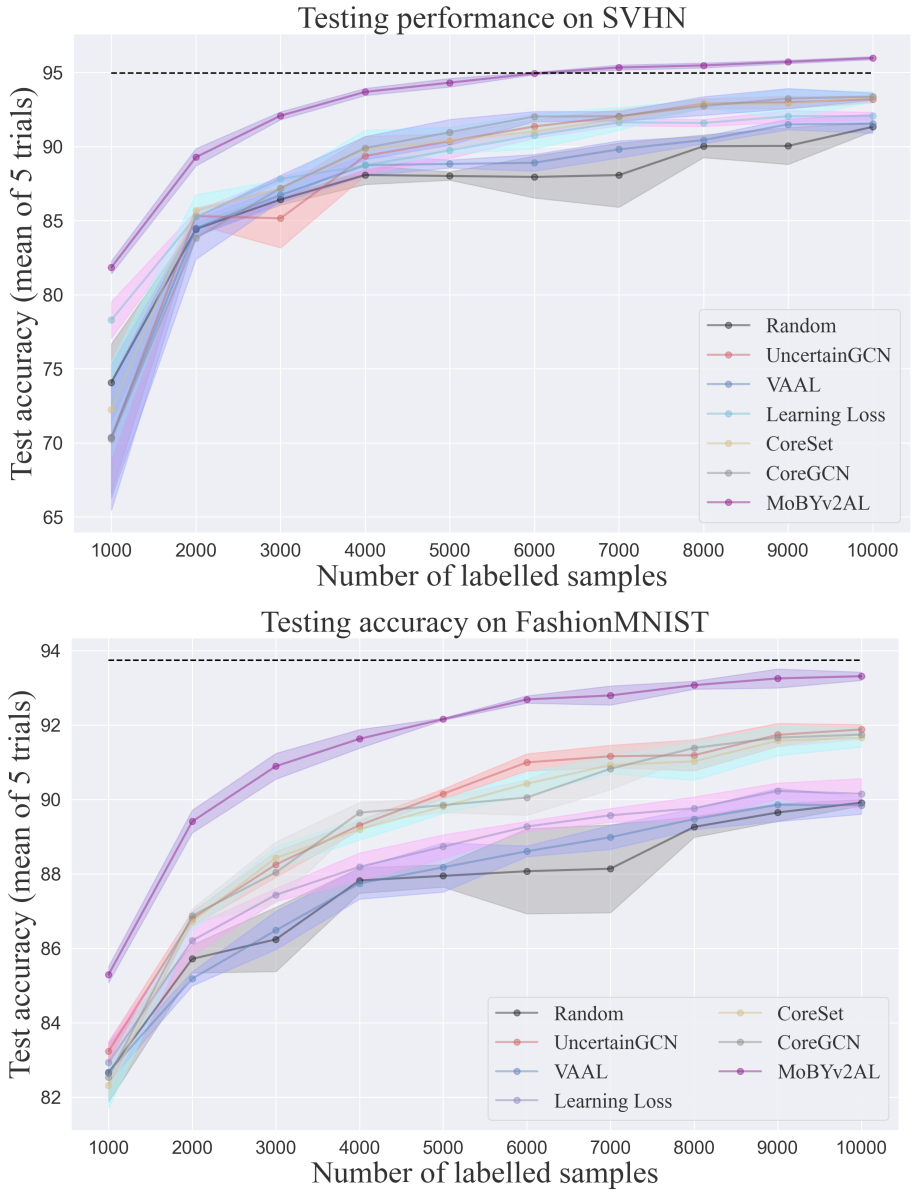

Figure 5.3: Evaluations on SVHN (Top), FashionMNIST (Bottom)

during real-world acquisition scenarios. Therefore, an imbalanced CIFAR-10 unlabelled set is simulated. Each of the ten classes has originally 5000 training examples. 5 of the classes are reduced to 500 images (resulting in a pool of 27500). The learner contains a ResNet-18 encoder, and it is trained with an initial set of 1000 labelled examples. MoBYv2AL is applied together with the other baselines for 7 cycles. Figure 5.4(top) presents the ability of MoBYv2AL to outperform the previous methods even in possible real-world environments. Investigation of long-tail distributions is still part of future work.

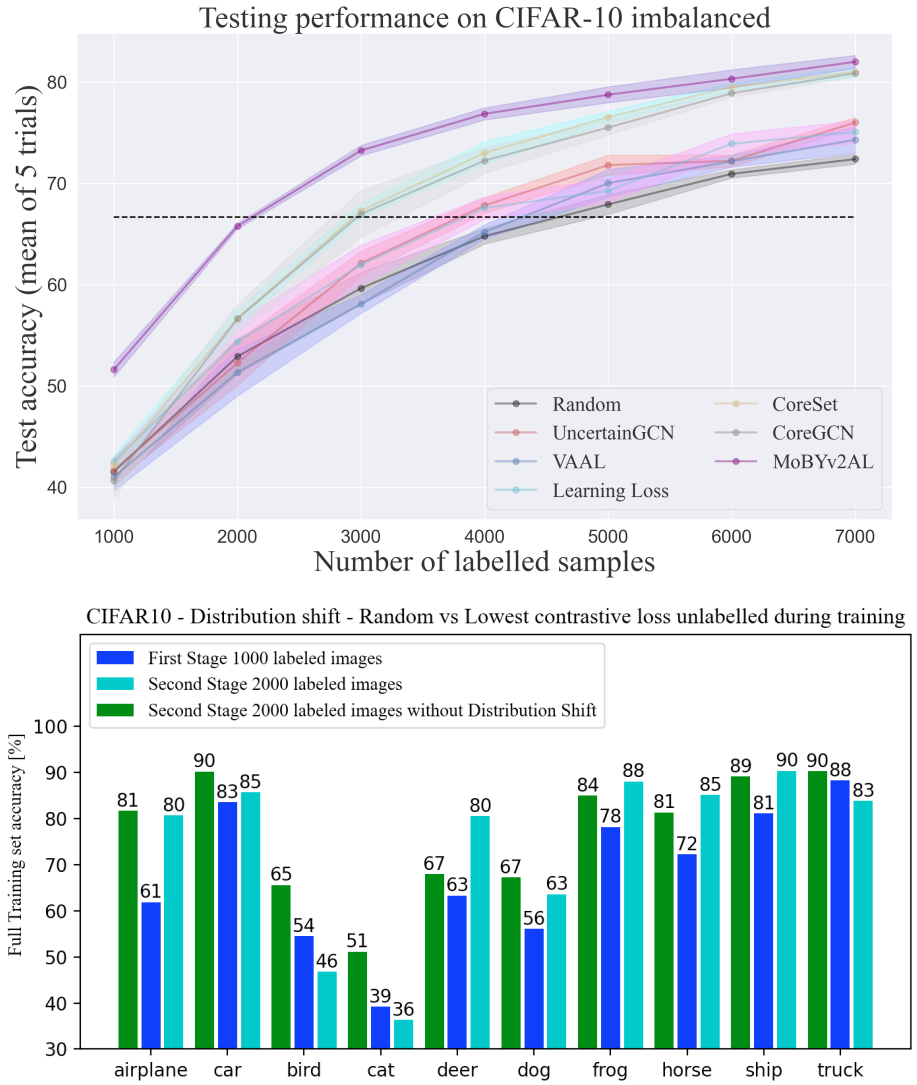

Figure 5.4: CIFAR-10 imbalanced dataset experiment(Top); Mitigating the distribution shift with MoBYv2(Bottom)

5.4.3 *Distribution shift discussion*

In DAL, the cyclical process of re-training the learner with the new labelled data may result in optimising to different local minima. Therefore, the exploration and exploitation of the AL method will be affected by this distribution shift at every stage. During experiments, this is commonly shown through jaggy curves (especially for uncertainty-based methods like MC Dropout (Gorriz et al., 2017), DBAL (Gal and Ghahramani, 2016) or UncertainGCN. To address this known issue (Kirsch et al., 2021), MoBYv2AL is analysed on the entire CIFAR-10 training set when providing 1000 and 2000 samples.

The dark blue bars of each class in Figure 5.4 (bottom) level the corresponding classification accuracy with the first 1000 random samples. Tracking the performance of the entire set challenges the learner to prefer certain classes. Another set of images is selected with MoBYv2AL. Consequently, the resulting accuracy is displayed by the cyan bar.

It can be clearly observed that the minima shifted in a different direction, where only some classes improved at the expense of the others. To mitigate this shift, this Section investigates what impact the unlabelled samples have in the end-to-end semi-supervised training. These samples play a key role in building up the dictionary of keys. The first insight is that the CoreSet selection on MoBYv2 data representation targets primarily high contrastive samples. This effect can be controlled by customising the unlabelled set deployed in training the learners.

To this extent, the proposed solution is to use the unlabelled data with the lowest contrastive loss. In 5.4 (bottom), with green bars, the performance of this mechanism is displayed. From an initial 1000 set accuracy (dark blue), there is an effective linear increase for all ten classes. This behaviour is consistent throughout all the previous quantitative experiments as well.

5.4.4 *SSL modules variation and Ablation Study*

The proposed design of MoBYv2AL is investigated with a set of ablation experiments and by varying its SSL module. Table 5.2 swaps in the end-to-end training pipeline the original version of MoBY (Xie et al., 2021) and the preceding SSL state-of-the-arts, MoCov2 (Chen, Fan, Girshick and He, 2020) and BYOL (Grill et al., 2020). Apart from MoBY, the learner did not converge on any selection cycle with the other SSL modules. Thus, the setup of large batches and specific training conditions (low learning rates, cosine scheduler) and learners can hardly adapt to this semi-supervision configuration.

For MoBYv2, the weak-augmented inferences to the learner stabilise its performance in regard to the original version. Furthermore, this method distances by 4% class accuracy with each AL cycle.

Table 5.2: Variation of SSL pipeline. Average testing performance (5 trials) on CIFAR-10 for 3 AL cycles with ResNet-18 encoder

SSLmodel / No. of labelled	1000	2000	3000
MoCov2	11.62±.9	11.92±.6	12.89±.6
BYOL	12.32±.7	11.72±.4	11.47±.2
MoBY	62.62±.4	72±.5	76.43±.1
MoBYv2	**63.06±.5**	**76.04±.6**	**80.63±.3**

Table 5.3: Ablation study of MoBYv2. Average testing performance (5 trials) on CIFAR-10 for 3 AL cycles with ResNet-18 encoder

MoBYv2 / No. of labelled	1000	2000	3000
w/o Discriminator	60.44±.4	72.53±.8	77.89±.3
w/o MLP Projector	58.57±.6	71.96±.5	77.02±.6
w/o Strong Augmentation	47.7±.4	58±.5	64.85±.5
MoBYv2	**63.06±.5**	**76.04±.6**	**80.63±.3**

One can argue that the SSL framework comprises several building blocks, and its implementation can deter developers. While there is significant dominance of MoBYv2 in AL selection, the motivation is continued by the relevance of each part in Table 5.3. In the ablation evaluation, the queue Discriminator and the MLP projectors are removed. As a result, a continuous accuracy drop is noticed. Projecting more prominent features and simulating the asymmetry brings the advantage of contrastive learning in MoBYv2. Moreover, strong augmentations also play a crucial role in the SSL pipeline.

5.4.5 *Selection function analysis*

The proposed pipeline, MoBYv2AL, can easily adapt to multiple selection methods. Here, the choice of CoreSet from section 5.3.2 is quantitatively motivated. Therefore, MoBYv2 is re-evaluated on CIFAR-10/100 benchmarks in Figure 5.5. The selection of the new budget is varied between random, maximum class entropy and CoreSet. Intuitively, the effect of selecting unlabelled examples with high contrastive loss is also analysed.

In both benchmarks, sampling with random or max entropy benefits the less MoBYv2AL pipeline. On the other hand, a representativeness-oriented

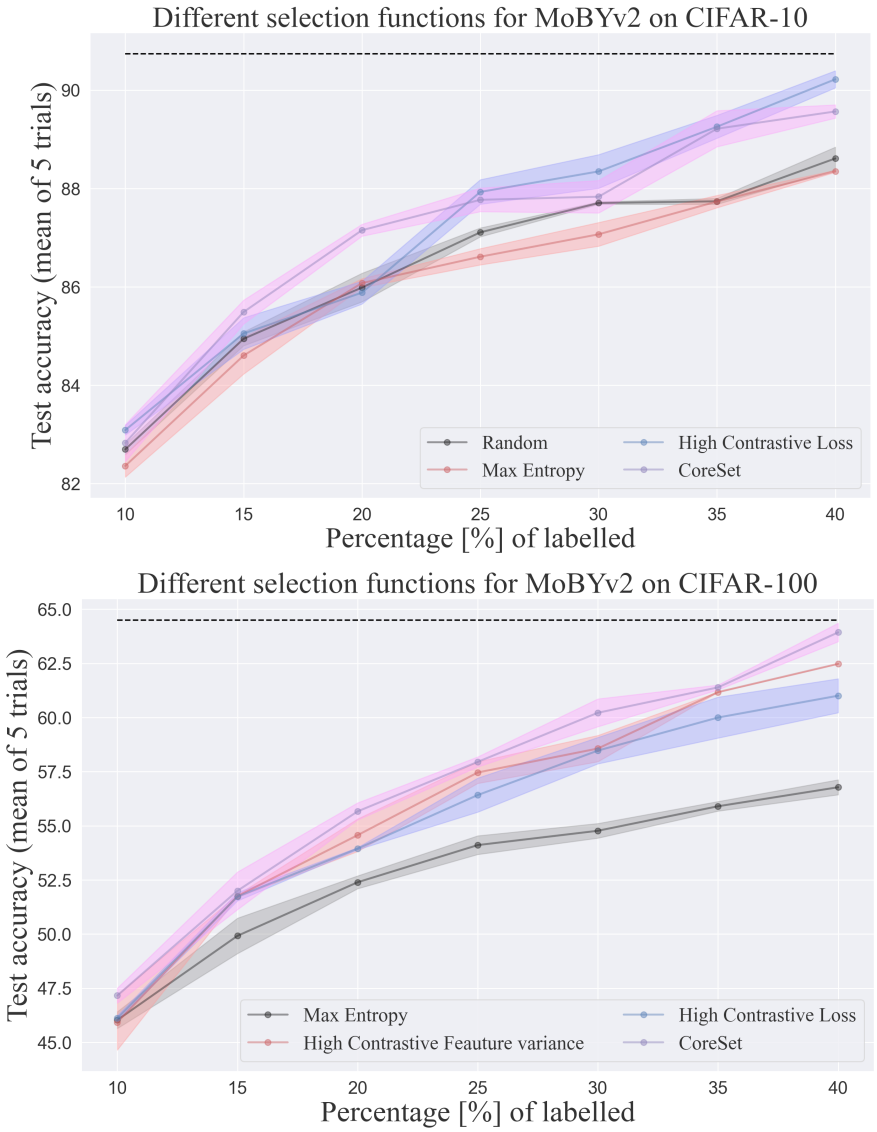

Figure 5.5: Quantitative evaluation with different selection functions for CIFAR-10 (Top), CIFAR-100 (Bottom)

method like CoreSet suits the hypothesis better. When sampling with high contrastive loss, repetitive examples were detected from some specific classes. This can be explained by higher contextual variance in that category. Specifically, on CIFAR-10, animal classes (cat, deer, dog), with stronger patterns, were preferred than the vehicle ones (car, truck, ship).

For a better visual analysis, a toy-set experiment was simulated with the first five classes from SVHN. Here, t-SNE (van der Maaten and Hinton, 2008) representations are taken from the MoBYv2AL query encoder outputs of unlabelled data. In Figure 5.6, the samples marked with crosses construct the new labelled set. The selection behaviour of the Max Entropy and CoreSet can

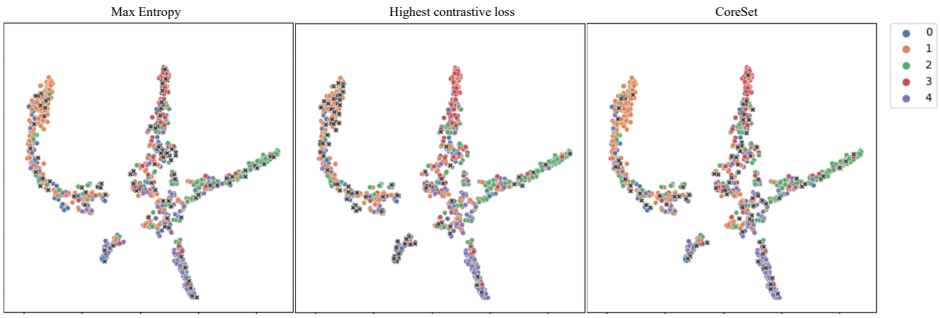

Figure 5.6: Qualitative AL selection analysis on MoBYv2. t-SNE representations at the first selection stage for five classes of SVHN

be interpreted as expected: on the left side, the uncertainty-based technique tracks the most class-variant images; CoreSet, on the right side, samples both in and out-of-distribution according to the Euclidean space.

Table 5.4: Multi-stage SSL-AL vs Jointly end-task AL.Testing performance on CIFAR-10 with ResNet-18 encoder

MoBYv2	1000	2000	3000
Multi-stage semi-supervised	34.8±.1	34.96±.2	35.09±.1
Jointly with end-task	63.06±.5	76.04±.6	80.63±.3

Table 5.5: Semi-supervised learning comparison. Testing performance on CIFAR-10 with ResNet-18 encoder

SSL method	Supervised	MoCov2	BYOL	DINO	MoBYv2
CIFAR-10 Test Acc	90.08	76.7	77.89	81.2	**88.62**

5.4.6 *SSL results and multi-stage AL*

MoBYv2 SSL for the AL strategy is designed in a joint manner with the end task. Despite this, the recent work Bengar et al. (2021) that proposes contrastive learning with SimSiamChen and He (2021) adopts multi-stage learning for the learner. The pipeline proposed fails to sample better than random in the AL paradigm. In Table 5.4, MoBYv2 is tested with the multi-stage training (self-supervised contrastive learning and second task fine-tuning) for CIFAR-10. It can be observed that the performance suffers in the context of the end task, where limited labelled examples are used. Similarly to Bengar et al. (2021), a minor improvement is also noticed when adding more selected data with CoreSet.

To this extent, in Table 5.5 the entire training set is used during fine-tuning. This experiment is re-iterated for SSL cross-validation with MoCov2 (Chen, Fan, Girshick and He, 2020), DINO (Caron et al., 2021) and BYOL (Grill et al., 2020). Even in this multi-stage setting, MoBYv2 is superior to the other state-of-the-art SSL.

5.5 CONCLUSIONS, LIMITATIONS AND FUTURE WORK

In this chapter, it was presented an SSL model-based AL framework for image classification. The main contributions lie in the task-aware contrastive learning pipeline. MoBYv2AL retains the higher visual concepts and aligns them with the downstream task. The end-to-end training is efficient and modular, allowing diverse learners and sampling functions.

Through the experiments conducted, it was shown that the SSL could mitigate the 'cold-start' problem by leveraging the visual feature information from the unlabelled samples as well. Under this configuration, a solution for the distribution shift effect was proposed by enforcing the unlabelled samples with the lowest contrastive loss in training the SSL.

It was demonstrated with a clear gap in state-of-the-art results on four datasets. The method showed robustness even in simulated class-imbalanced data pools. Mimicking the real-world scenarios, MoBYv2AL should be tested in the long-tail distribution scenario for future work.

While this method has been successful for the image classification task, its modular capability can easily be transferred to other tasks. This would also be part of future work, however, some limitations could be encountered. One of the limitations is firstly in regard to the data augmentation process in the case where the input data are not RGB/greyscale images. Another limitation some may consider is the computational and hardware complexity of doubling the initial learner architecture.

In the context of the targeted task of this thesis, 3D-HPE, a thorough investigation of the data augmentation should be established before adapting MoBYv2AL. One of the pioneering self-supervision works in this direction is presented for monocular RGB 3D-HPE by Spurr et al. (2021). Here, the authors define the negative and positive examples according to the articulation space and weak augmentations, respectively.

Finally, part of the future work will also have to investigate the saturation nature that MoBYv2AL may have at the initial training stage. Uncertainty-based AL selection techniques applied adaptively may balance the exploration-exploitation trade-off in the saturation case.

6

MULTI-MODAL VARIATIONAL AUTO-ENCODER FOR ACTIVE LEARNING

6.1 COURSE OF OBJECTIVES WITH JUSTIFICATION

This chapter extends the 3D-HPE task to multi-modal inputs, which is a currently popular topic. Keeping the tackled limitations and converging the previous AL insights, the final question is addressed:

> ""How do these modalities influence the performance gain in the context of DAL?"

In the attempt to answer the question, the upcoming sections bring back the 3D-HPE task as primarily for AL. From Chapter 3, it was denoted that architectural constructs like Bayesian approximations may not be efficiently feasible for sampling. Although it can provide meaningful uncertainties and competitive AL selection function like CKE, parametric samplers are still less intrusive and easier to be deployed like in Chapter 4 and Chapter 5. Therefore, this chapter follows the same trend of DAL, and it proposes a VAE-based generative sampler for multi-modal AL.

Since the DL expansion to 3D-HPE, depth-based datasets (Tang et al., 2014; Tompson et al., 2014; Yuan, Ye, Stenger, Jain and Kim, 2017) have been tackled in the beginning while providing the most performing models with accurate and realistic hand articulations (Chang et al., 2018; Oberweger and Lepetit, 2017; Xiong et al., 2019; Ye et al., 2016). It was observed that as these models attained averaged MSE of 6-7 mm and even fitting to the annotation noise (Yuan, Ye, Garcia-Hernando and Kim, 2017*b*), other researchers from the 3D-HPE community have started to tackle once again single RGB inputs. This is due to the capacity of DL models to escalate the highly non-linear function in pose reconstruction. The majority of the state-of-the-art (Baek et al., 2019; Ge et al., 2019; Spurr et al., 2018; Yang et al., 2019) on this direction have integrated at least another modality to compensate for the limitations of the RGB input (in terms of illumination variation, background environment and occlusion). This trend continuously motivates the proposed interest in AL together with the rise of powerful cross-domain DL architectures like CLIP (Radford et al., 2021) and Perceiver (Jaegle et al., 2021).

Another motivation for promoting the multi-modal approach stands on the availability of new 3D-HPE datasets that provide at least two modalities. From these, the most well-known are STB (Zhang et al., 2017), RHD (Zimmermann and Brox, 2017), FreiHand (Zimmermann et al., 2019) and InterHand2.6M (Moon et al., 2020). In order to keep the AL selection research efficient, only the first three datasets are considered in the experiment section. InterHand2.6M as in BigHand2.2M (Yuan, Ye, Stenger, Jain and Kim, 2017) is formed from continuous large sequences with 30 frames per second. This would slow the runtime of the experiments, and the issue of sample redundancy would still have to be addressed.

Moving to the VAE-based AL sampler, its concepts are inspired from VAAL and Chapter 5. MoBYv2AL makes use of the SSL to improve the visual concepts by leveraging the unlabelled, whereas the multi-modal sampler aims to bring a better latent representation than the learner with a VAE. Distinctively from VAAL, however, the generative framework is not trained in an adversarial manner but with a combination of modality reconstructions and 3D-HPE regression objective. The poor performance for AL selection of VAAL was denoted in Chapter 4 due to two reasons: one is linked to the lack of representativeness between the sampler and the learner; the second consists of the uncertainty that the adversarial discriminator creates between labelled and unlabelled while making it unaligned to the one from the learner.

With the above motivations and mentioned prior issues, this chapter proposes **MVAAL**, a **M**ulti-modal **V**ariational **A**uto-encoder AL sampler, where multiple modalities are trained for an aligned VAE latent space similarly to AlignMM (Yang et al., 2019). The reason for aligning the modalities is to provide a shared space from which the selection function can provide informative sets of images to the learner. Moreover, the objective of MVAAL is aligned with the learner so that the latent space is skewed towards the end task. To validate this proposal, AL selection is applied with various samplers on the aligned latent space obtained with the PoE approximation. Although this work is part of ongoing research, promising results are shown in the experiment section for all the multi-modal datasets STB, RHD and FreiHand.

Summarising the next sections, related works 6.2 are discussed within the multi-modal AL and 3D-HPE research areas, followed by the proposed methodology and the experimental section. The methodology will shortly present the structure of the learner 6.3.1 and the MVAAL sampler 6.3.2 with each selection criteria in detail. The experiment Section 6.4 will qualitatively analyse the reconstructed hand poses from MVAAL, and it will conclude with the numerical AL evaluation on the three multi-modal 3D-HPE datasets.

6.2.1 *Multi-modal 3D-HPE*

The initial works that combined RGB with other modalities for 3D-HPE appeared together with the proposed datasets STB, RHD and FreiHand. However, monocular RGB state-of-the-art has just recently outperformed these benchmarks with the algorithms from Ge et al. (2019) and Baek et al. (2019). Both algorithms use 3D mesh reconstruction and masking to improve and transform the 2D pose estimation from RGB. Apart from some architectural choices, the key difference stands in the way of creating the 3D hand mesh. Baek et al. (2019) proposed to use the MANO synthetic model (Romero et al., 2017) to infer the articulation, while Ge et al. (2019) claims to generate more realistic synthetic hand model with GCN.

Noting that these pose estimators integrate the second modality as a functional part of their network design, MVAAL looks for learners with multiple inputs design. The works that leveraged depth information or other inputs and used generative networks between these modalities are Cross-Modal (Spurr et al., 2018), Depth-Regulariser (Cai et al., 2018) and AlignMM (Yang et al., 2019). Depth-Regulariser consists actually of two separate networks with a shared hand pose layer. The first network does pose estimation from RGB, and the second generates depth maps from the pose layer. The shared layer, however, still requires pose annotations. To overcome this constraint, Cross-Modal creates the shared latent space by alternating between RGB and depth modalities. This alternation might be ineffective, and the shared latent space can get misaligned. Incrementally, AlignMM addresses this issue and proposes an alignment with PoE for a unified and generalised VAE framework that can support more than two modalities. This motivated the derivation of MVAAL's strategy and sampling from AlignMM (Yang et al., 2019).

6.2.2 *Multi-modal AL*

AL for multi-modality has barely been explored over the past years. Although multi-modal data starts to be widely available to multiple areas of research from automotive (Haussmann et al., 2020) to medical (Budd et al., 2021), there are still concerns about how optimal this data is in relation to the new DL architectures like residual CNNs (He et al., 2016; Zagoruyko and Komodakis, 2016) or transformers (Dosovitskiy et al., 2020; Wu et al., 2020). The first

multi-modal AL framework was proposed for autism therapy in child-robot interactions (Rudovic et al., 2019). This application shows how essential AL can be when engagement estimation has to be transferred to other children so that more personalised data is attributed accordingly. Modalities are gathered from pose networks, sounds and watch accelerometers for majority voting of the engagement. The AL framework trains a Reinforcement Learning (RL) policy with a Q-function to provide the set of modalities with the highest engagement for labelling.

Still in the medical field, the only other work that treats multi-modality for AL is MMFN-AL (Gao et al., 2021). This method is applied for liver fibrosis diagnosis on four modalities, one ultrasound and three shear wave elastography. The AL sampler is task-aware and uses the identical feature encoders as in the learner, but compared to MVAAL, it just concatenates the latent space from all the modalities. In terms of selection criteria, random sampling is used against the uncertainty-based methods of highest entropy (Shannon, 1948) and from MC Dropout (Gorriz et al., 2017). Despite the lack of multi-modal integration in the latent space, MMFN-AL shows distinctive performance with MC Dropout against random, but for single modality reference. The main concerns can be observed when comparing selections in the multi-modal experiment. It is observed that the AL function, MC Dropout, performs as well as random. Therefore several query functions will be approached and tested with MVAAL in the following sections.

6.3 METHODOLOGY

The proposed multi-modal AL pipeline for 3D-HPE is presented in detail here. Firstly, the learner's architecture is discussed, which follows a multi-modal version of DeepPrior++ (Oberweger and Lepetit, 2017). Secondly, where the main contribution lies, the MVAAL sampler is theoretically explained together with its corresponding query functions.

6.3.1 *Multi-modal 3D-HPE learner*

For 3D-HPE, the DL solutions can broadly be divided between generative and discriminative. On the one hand, the generative models assume to have an architecture that commonly reconstructs heat maps with normal distribution intensity around the joint locations. On the other hand, discriminative models break down the high-order representations by regressing directly to joint coordinates. There are many advantages and disadvantages to both categories,

however, for simplicity, and to ease the computational runtime, the network architecture of the learner is chosen to be discriminative.

The hand pose estimator is designed following the DeepPrior++ architecture. Thus, the feature encoder adopts a ResNet-18 (He et al., 2016) design. Although the authors of DeepPrior++ proposed a deeper residual network, ResNet-50, it was experimentally noticed that the performance is marginal compared to the ResNet-18 backbone.

Therefore, given m modalities, the learner is composed of m ResNet-18 feature encoders, each expressed as a function $f_m(\mathbf{x}_m)$ ($\mathbf{x}_m$ the input modality m). To fuse the m latent spaces, a concatenation is applied before regressing to the 3D joint coordinates. The regression is mapped by a discriminator with three fully-connected layers.

Before moving to the AL sampler, it is important to discuss the modalities and the pre-processing used for the three datasets, STB, RHD and FreiHand. For clarity and uniformity within the datasets, only $m = 2$ modalities are used from each dataset. Specifically, the modalities applied from STB and RHD are colour and depth, while in the FreiHand case, the inputs are colour and segmentation maps. Other modalities like heat maps and point clouds can also be extended in this pipeline (as in AlignMM (Yang et al., 2019)).

To assist the 3D-HPE several pre-processing steps are applied to both inputs and annotations. Firstly, the 3D coordinates are normalised, making the hand input scale invariant. Moreover, training to a normalised number offers better convergence stability over the discriminator's parameters. Secondly, it was shown in prior works (Oberweger and Lepetit, 2017; Ye et al., 2016) that having the palm root key-point as a reference makes the estimation translation invariant.

Instead of deploying a hand detector, a centre crop is applied to the images according to the key point ground truths. Apart from this, a normalisation of the pixel values and a re-scaling to 128×128 are the final transforms to the modalities.

6.3.2 *MVAAL*

The aim of the MVAAL sampler, similarly to MoBYv2AL, is to diversify the data representations with features from both labelled and unlabelled data. This has shown to be effective in increasing the representativeness of the learner, an essential requirement to AL sampling. In order to achieve this, VAE architectures are chosen instead of a contrastive learning siamese pipeline (as in MoBYv2AL). The motivation is derived from the successful

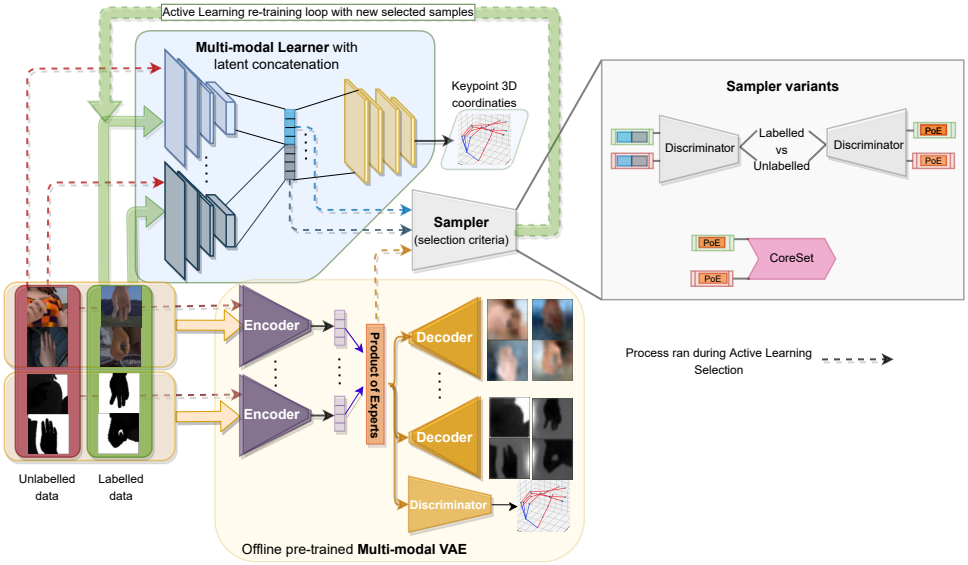

Figure 6.1: The proposed AL pipeline: the multi-modal 3D-HPE learner (Top); MVAAL sampler (Bottom) with selection functions (Right); Two modalities are represented in the figure, but more can be added. MVAAL constitutes of the pre-trained multi VAE with the PoE alignment; the pose discriminator and the selection function (as sampler variants). MVAAL fine-tunes the pose discriminator with the new labelled subset that was assigned to the learner after each AL cycle.

3D-HPE arts of Cross-Modal (Spurr et al., 2018) and AlignMM (Yang et al., 2019) where the VAE architectures have increased the performance in the cross-modal setting. Another advantage of the VAE, in the context of AL is that the training can be done on the whole pool of data as no labels are required. Therefore, the training can be efficiently done offline, out of the AL cycle. Below, MVAAL will be detailed with the task-aware discriminator and its selection methods.

6.3.2.1 *Aligned latent space for multiple modalities*

Given $m = 2$ modalities $\mathbf{x}_1, \mathbf{x}_2$, the multi-modal VAE follows the encoder architecture of the learner with $\phi_{\mathbf{x}_1}$ and $\phi_{\mathbf{x}_2}$ parameters. The pre-processed inputs can be encoded by $q_{\phi_{\mathbf{x}_1}}$ and $q_{\phi_{\mathbf{x}_2}}$ to the latent spaces $\mathbf{z}_{\mathbf{x}_1}, \mathbf{z}_{\mathbf{x}_2}$. For reconstructing the images, the architectures of the encoders are mirrored to two decoders with the parameters $\theta_{\mathbf{y}_1}, \theta_{\mathbf{y}_2}$, respectively (where $\mathbf{y}_1$ and $\mathbf{y}_2$ are the corresponding target modalities).

These two VAEs can be treated separately, unidirectional, and trained by maximising the ELBO:

$$\begin{aligned}\mathcal{L}_{\text{ELBO}}(\mathbf{x}_1; \mathbf{y}_1; \theta_{\mathbf{y}_1}, \phi_{\mathbf{x}_1}) = \mathbb{E}_{z\sim q_\phi} \log p_{\theta_{\mathbf{y}_1}}(\mathbf{y}_1|\mathbf{z}_{\mathbf{x}_1}) - \\ \beta \text{KLD}(q_{\phi_{\mathbf{x}_1}}(\mathbf{z}_{\mathbf{x}_1}|\mathbf{x}_1) \| p(\mathbf{z}_{\mathbf{x}_1})),\end{aligned} \tag{6.1}$$

for the first modality, where KLD is the KL divergence, β is a latent space reconstruction parameter (Higgins et al., 2017) and $p(\mathbf{z}_{\mathbf{x}_1})$ is the Gaussian prior on the this space. However, the goal is to align the two modalities under a shared latent space while being able to generate from one modality to another. Following AlingMM, MVAAL extends the PoE alignment proposal to multiple separate decoders. The generative nature of these decoders results in a relation between the posteriors of both joint and single modalities. In Wu and Goodman (2018), it was shown that the joint posterior is proportional to the product of individual ones $q(\mathbf{z}_{\text{joint}}|\mathbf{x}_1, \mathbf{x}_2) \propto p(\mathbf{z}_{\text{joint}}).q(\mathbf{z}_{\text{joint}}|\mathbf{x}_1)q(\mathbf{z}_{\text{joint}}|\mathbf{x}_2)$. This has been obtained by approximating the true posteriors for each modality while containing them in the family of the variational counterparts.

The joint objective with the product of two Gaussian experts (PoE) $G^*(\mathbf{z}_{x_1}, \mathbf{z}_{x_2})$ is:

$$\begin{aligned}\mathcal{L}_{VAE}(\phi_{x_2}, \phi_{x_1}, \theta_{y_1}, \theta_{y_2}) = E_{\mathbf{z}_{joint} \sim G^*(\mathbf{z}_{x_1}, \mathbf{z}_{x_2})} \log p_{\theta_{y_1}, \theta_{y_2}}(\mathbf{y}_1, \mathbf{y}_2 | \mathbf{z}_{joint}) + \\ E_{\mathbf{z}_{x_1} \sim q_{\phi_{x_1}}} \log p_{\theta_{y_1}, \theta_{y_2}}(\mathbf{y}_1, \mathbf{y}_2 | \mathbf{z}_{x_1}) E_{\mathbf{z}_{x_2} \sim q_{\phi_{x_2}}} \log p_{\theta_{y_1}, \theta_{y_2}}(\mathbf{y}_1, \mathbf{y}_2 | \mathbf{z}_{x_2}) - \\ \beta(KLD(q_{\phi_{x_1}}(\mathbf{z}_{x_1} | \mathbf{x}_1) \| p(\mathbf{z})) + KLD(q_{\phi_{x_2}}(\mathbf{z}_{x_2} | \mathbf{x}_2) \| p(\mathbf{z}))).\end{aligned} \tag{6.2}$$

The product of the two Gaussian experts results in a Gaussian as well. Therefore, for the two means μ_1, μ_2 and covariances Σ_1, Σ_2, (of the two modalities) the resulted joint mean and covariance are:

$$\mu_{joint} = (\frac{\mu_1}{\Sigma_1} + \frac{\mu_2}{\Sigma_2}) / (\frac{1}{\Sigma_1} + \frac{1}{\Sigma_2}) \tag{6.3}$$

$$\Sigma_{joint} = (\frac{\Sigma_1 \Sigma_2}{\Sigma_1 + \Sigma_2}) \tag{6.4}$$

The final module of MVAAL is the pose discriminator integration in the loss function $\mathcal{L}_{VAE}$ so that the representations are directed to the joint reconstruction. The pose discriminator architecture replicates the one from the learner. The inputs for this module are the joint means, μ^L_{joint}, of the annotated data as its loss minimises with l2 norm of the 3D hand coordinates.

In terms of training, as mentioned before, the multi-modal VAE is optimised once and independent from the learner. To keep MVAAL task-aware, however, with every AL cycle and new annotated subset, the pose discriminator is trained. Thus, it automatically fine-tunes the VAE at the same time. The common objective can be expressed as follows:

$$\mathcal{L}_{MVAAL} = \mathcal{L}_{VAE} + \mathcal{L}^2_{pose}. \tag{6.5}$$

This practice has also proven effective in the MoBYv2AL pipeline.

6.3.2.2 *Selection methods for MVAAL sampler*

Once MVAAL is trained with the task discriminator, the AL selection of unlabelled data can start. Here, two query functions are analysed, one that tracks uncertainties and another that uses geometric distances.

Similarly to the GCN sampler proposal in Chapter 4, the uncertainty-based selection deploys a binary discriminator to distinguish between the PoE representations from labelled and unlabelled data. This is composed of two fully-connected layers and takes as inputs equal batches of labelled and

unlabelled representations from MVAAL. In the end, the trained uncertainty discriminator ranks the most uncertain unlabelled samples as UncertainGCN. The geometric selection deploys the CoreSet method as in all the previous AL frameworks (CKE, CoreGCN and MoBYv2AL) again. Both these selections have shown significant improvements to various datasets, further motivating their integration into this work.

6.4 EXPERIMENTS

6.4.1 *Quantitative experiments*

MVAAL is evaluated quantitatively in a pool-based DAL scenario. As in previous chapters, the learner starts training with an initial pool of annotated examples. Then, AL is applied, in this case, MVAAL training and selection, to the unlabelled pool so that a new fixed batch (by a budget) is annotated. In preparation for the experiments on the three multi-modal datasets, STB, RHD, and FreiHand, some details regarding the training and the AL settings will enfold.

6.4.1.1 *Learner and MVAAL training settings*

The 3D-HPE learner is trained for 100 epochs with a batch of 32, and it is optimised with Adam (following a learning rate scheduler from 0.0001 and reducing by ten times at 40 and 60 epochs). The regression minimises the normalised 3D coordinates with MSE loss.

MVAAL has its initial pre-training on 140 epochs, and the batch size is equal to the learner. It is also optimised in the same way, but the learning rate starts at 0.001, and it has a scheduling protocol when the loss plateaus. Specifically, to the VAE, the β parameter is set at 0.5 in Equation 6.2. For the fine-tuning of the pose discriminator, the settings are kept the same, however, the number of epochs is reduced to 60 (the combined loss converges rapidly on the subsets of new annotated data).

6.4.1.2 *AL settings and Baselines*

AL deploys the pool-based scenario where each training set is considered unlabelled. Thus, the three pools of data are 15,000, 41,258 and 32,560 for STB, RHD and FreiHand, respectively. During the experiments, it was noticed that all datasets present sample redundancies, therefore, before the selection process, random filtering is applied to create subsets. From these, the AL

selection is effectively made. The reduced subsets are 7,000, 10,000 and 8,000 (for the corresponding dataset STB, RHD and FreiHand). The exploration is performed over five AL cycles with a budget equal to the initial training set of 10% from each dataset. The experiments are conducted over five trials to offer a stable averaged MSE of the 3D joint coordinates.

In terms of baselines, apart from random sampling and the two selection functions on MVAAL, ranking the uncertainty on the discriminator trained with features from the learner is also investigated. Moreover, the standard CoreSet is added as a known competitive baseline. Both these baselines demonstrate, in an ablative way, the motivation of using MVAAL with the PoE.

6.4.1.3 *AL quantitative evaluations*

Initially, the multi-modal learner is trained on every dataset using the entire sets. Measured in mm, the averaged MSE obtained are 12.19 for STB, 15.25 on RHD, and 40.38 mm for FreiHand. As a confirmation that the multi-modal pipelines benefit the second modality, an RGB-based learner was trained on RHD. The error increased to 19.59 mm in this case, attesting to the multi-modal hypothesis.

Moving to the multi-modal experiments with MVAAL, in Table 6.1, the STB dataset is evaluated on five cycles. From all three datasets, STB has already been saturated in testing performance by state-of-the-art hand pose estimator (Baek et al., 2019; Ge et al., 2019). It can be noticed that random sampling and uncertainty sampling when the discriminator is trained on the concatenated learner's features, reaches an error deviation of 3 mm at 50% annotated images. All the other AL query functions fail to provide better informativeness than a simple random selection.

Table 6.1: Quantitative AL results on STB dataset. Averaged joints MSE over 5 trials. % of annotated data

AL selection method	10%	20%	30%	40%	50%
Random sampling	21.68	21.69	19.89	18.91	15.9
Feature Uncertainty Discriminator	21.68	20.38	19.62	17.85	15.62
PoE Uncertainty Discriminator	21.68	19.72	19.93	19.38	17.39
CoreSet	21.68	19.06	18.99	17.76	15.76
CoreSet on PoE	21.68	20.1775	20.67	17.77	16.84

Conversely, RHD and FreiHand present more challenging articulations and viewpoints, making it harder to reach lower errors. Tables 6.2 and 6.3 show the obtained results under the mentioned settings. For both RHD and FreiHand, it can be noticed that the lowest errors are from the selected datasets with the PoE discriminator. At 50% annotated dataset, MVAAL attains MSE with differences of 3 mm and 5 mm for RHD and FreiHand, respectively. The other selection methods yield relatively close results. On FreiHand, the gap is significant enough in regard to random. However, on the RHD dataset, the gain of AL sampling is still debatable, as in the STB experiment.

Table 6.2: Quantitative AL results on RHD dataset. Averaged joints MSE over 5 trials. % of annotated data

AL selection method	10%	20%	30%	40%	50%
Random sampling	33.78	25.65	21.74	19.91	18.49
Feature Uncertainty Discriminator	33.78	25.55	21.48	19.91	18.495
PoE Uncertainty Discriminator	33.78	**25.39**	**21.63**	**19.9**	**18.3**
CoreSet	33.78	26.59	22.95	20.79	19.29
CoreSet on PoE	33.78	26.52	22.28	19.67	18.92

Table 6.3: Quantitative AL results on FreiHand dataset. Averaged joints MSE over 5 trials. % of annotated data

AL selection method	10%	20%	30%	40%	50%
Random sampling	59.99	54.72	50.61	48.75	46.7
Feature Uncertainty Discriminator	59.99	53.76	50.32	47.64	46.02
PoE Uncertainty Discriminator	59.99	**54.01**	**48.45**	**46.79**	**45.86**
CoreSet	59.99	53.38	49.73	46.91	**45.37**
CoreSet on PoE	59.99	53	50.75	47.95	**45.3**

Because the gains of the AL methods are not convincing enough, mainly for STB and RHD, the last experiment in Table 6.4 deploys MVAAL from two modalities to a 3D-HPE learner trained only with RGB images. The experiment is conducted on RHD, but only the PoE discriminator, CoreSet and random sampling are used for comparison. Here, an interesting trend is shown. CoreSet manages to find a better subset of data initially but flattens the performance at later stages (likely due to lack of diversity). The PoE

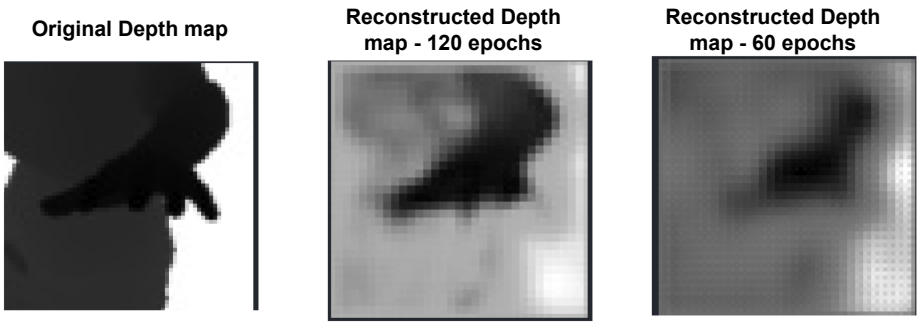

Figure 6.2: Depth-maps reconstructed with the multi-modal VAE at different training points.

discriminator has a poor selection in the beginning by generating highly uncertain poses. Despite this, with more cycles, it manages to reach a relatively low MSE with just a 2 mm deviation from the whole dataset training accuracy, 19.59 mm. In MMFN-AL (Gao et al., 2021), the methodology was proposed in a similar mono-modal setting, and a MC Dropout selection came to the same results and conclusions.

Table 6.4: Quantitative AL results on RHD dataset. RGB Single Modality. Averaged joints MSE over 5 trials. % of annotated data

AL selection method	10%	20%	30%	40%	50%
Random sampling	39.1	29.12	25.38	23.14	23.02
CoreSet	59.99	**28.26**	24.75	23.32	**24.08**
PoE Uncertainty Discriminator	59.99	29.76	**24.72**	**22.84**	**21.54**

6.4.2 *MVAAL Qualitative Analysis*

Qualitative observations were drawn to validate the pre-training of the multi-modal VAE. Although the combined loss converges smoothly and flattens after 50 epochs, it was noticed, according to Figure 6.2, that the reconstruction continues to improve. In this regard, keeping this qualitative evaluation throughout the training is recommended.

Figure 6.3 displays the original and reconstructed versions of four sets of RGB-D images from the RHD dataset. It can be observed that in the majority of cases, the images are reconstructed to a course level. Specifically, the background, hand shape and colour fit the originals. The depth map reconstruction between the two modalities is more robust due to the extra

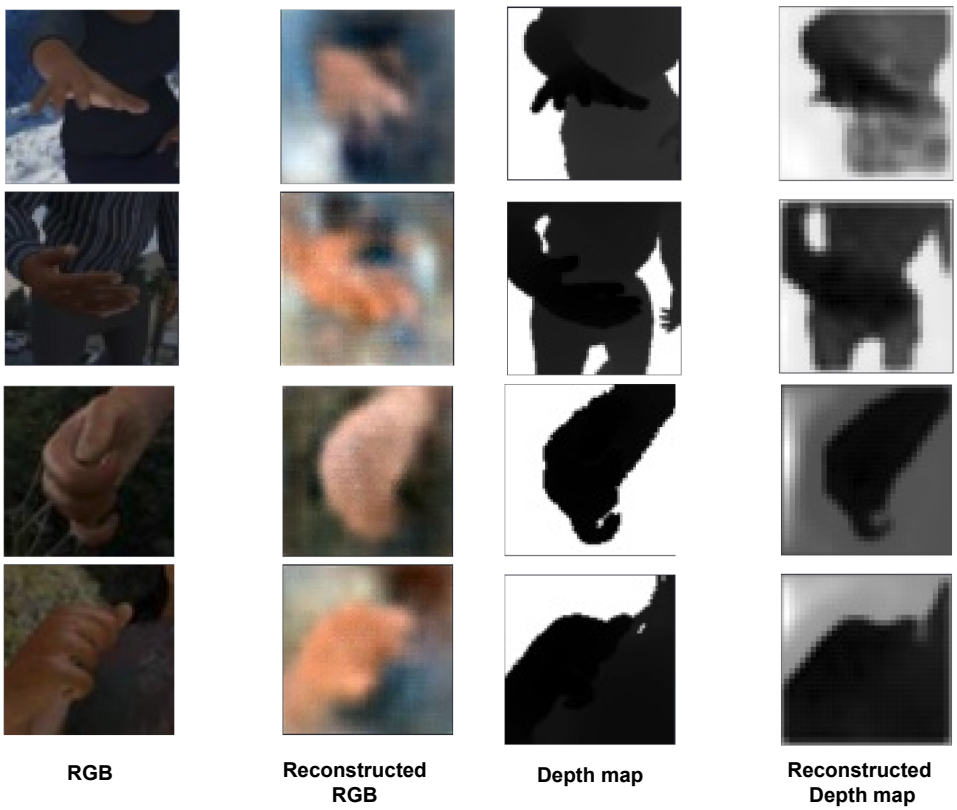

Figure 6.3: Multi-modal VAE reconstructed modalities from the PoE

two channels that the RGB decoder needs to upscale. Conclusively, the quality of the reconstructions suits the proposed pipeline while attesting to a meaningful, aligned latent space.

6.5 FUTURE WORK AND CONCLUSIONS

The proposed multi-modal AL method is still in a concept phase where future development would be required. This development will have to address the limited performance that multi-modal AL provides in regard to random sampling. The hypotheses of MVAAL would need to be tested and ablated against the other state-of-the-art AL function. This would apply only when an AL baseline would produce consistent performance gaps compared to random selection. Changes in the learner's pipeline and estimating the influence of each modality will be the upcoming parts of this research.

The presented SSL pipeline from chapter 5 and the GCN sampler from chapter 4 can be complementary modules to MVAAL. Therefore, the multi-modal learner can be designed according to MoBYv2AL, where the 3D-HPE regression loss is trained with the contrastive loss. As mentioned before, an

investigation of the different types of augmentation should be performed for each modality. As another possible training scheme, each modality can be seen as a negative example to each one if a separation between the modality domains is desired for the SSL. In terms of sampling, MVAAL may integrate the feature representations from MoBYv2AL (which are tackling the cold-start problem) and for future work, a new GCN sampler could be proposed.

Despite the limits in performance, MVAAL has shown to be a promising AL framework. The aligned latent space between all the modalities provides a unified representation to guide a better AL selection. As in MMFN-AL, the multi-modal AL pipeline with the uncertainty discriminator achieves a minimal error at 50% dataset when the learner uses only the RGB modality (with a 2 mm difference to training with the entire RHD dataset). With the high demand of DL architectures and multi-modal datasets, the approach of AL in optimising the data becomes even more relevant.

7

CONCLUSIONS AND FUTURE WORK

This chapter aims to gather all the contributions and achievements presented in the thesis. It also provides insights and the impact that it has on the AL research community. With a primary focus on 3D-HPE and other vision tasks, the methods presented achieve breakthrough results in sampling informative and customisable data for the DL architecture. Despite this, the encountered limitations and difficulties are discussed here as well. The chapter ends with a reflection on future work and a statement from the author.

7.1 CONTRIBUTIONS AND ACHIEVEMENTS OF THE THESIS

The overall goal of the thesis is to find efficient ways of sampling real data with AL. Apart from the methods themselves, the main contribution is that it pushes the research for the first time to the DL methods, exclusively to the 3D-HPE application. After a first attempt at solving the informativeness of the unlabelled data to the learning algorithm, the thesis diverges with a focus on the AL methodology as a universal solution. Therefore, dedicated samplers are investigated with trainable parameters to assist the learner in providing meaningful new data. The last part of the thesis integrates the previous methods and achievements back to 3D-HPE but in a multi-modal input form.

In Chapter 3, AL is approached for the first time to the 3D-HPE task where the learners are CNNs. To deal with the fact that the majority of the pose estimator is discriminative (regression-based), one of the contributions is the Bayesian approximation to these models so that confidence can be obtained with the posterior. The confidence measures are expressed as per-joint variances comprised of aleatoric (data-dependent) and epistemic (model-dependent) uncertainties. Through qualitative analysis, it was shown that the produced uncertainties are representative to the hand poses from all the datasets. Moreover, it investigated the ideal structure of the Bayesian DeepPrior learner that in the end, showed an improved overall pose estimation against the standard architecture. The second set of contributions is in regard to the AL method and sampling. The two uncertainties play a critical part in deploying the presented method, CKE. Thus, the geometric AL function

CoreSet includes during the cluster distance computation the corresponding epistemic uncertainties. As the main set of achievements, CKE has shown to produce lower average MSE on all three depth datasets than its direct comparative baseline, CoreSet. This is done in the context of AL where this performance is obtained with less amount of annotated subsets.

The Chapter 4 of the thesis approaches the AL problem from the sampler's perspective rather than the previous Bayesian approximation of the learner. Therefore, it proposes a novel task-agnostic parameterised method that takes advantage of higher order representations from a GCN. It was shown that the pool-based AL methodology could be applied to both classification and regression tasks in the context of DL. Two selection criteria have been studied, one that ranks uncertainty (UncertainGCN) and the other that integrates representativeness with CoreSet (CoreGCN). Both, through comprehensive experiments, have produced highly competitive results for over eight datasets, specifically, four for image classification, three for the 3D-HPE task and one for synthetic sampling.

Chapter 5 extends the GCN sampler and addresses some of the faced issues like the 'cold-start' problem and the distribution shift. The methodology presents an SSL model-based AL framework for image classification. It uses for the first time the contrastive learning insights to DAL. The main contributions lie in the pipeline MoBYv2AL, which leverages visual representation from all the data and aligns the objective with the downstream task. Following the same modular and parameterised trend of the GCN sampler, MoBYv2AL can also be used in other applications.

The AL quantitative experiments have shown that the proposed methodology pushes the performance with a large margin for all four datasets. Even from the initial AL cycle, leveraging the representations from the unlabelled helps to solve the 'cold-start' problem. For a smooth increment in performance with each labelled subset, a solution was proposed to tackle the distribution shift. Particularly, less contrastive unlabelled samples were promoted in the semi-supervised re-training with each AL cycle. Still on the performance, certain experiments like SVHN or the imbalanced CIFAR-10 have demonstrated that MoBYv2AL can surpass the full dataset training baseline.

Although the work in Chapter 6 is still in progress, certain insights and directions have been deducted. The main contribution, however, is in the approach and application of AL for multi-modal 3D-HPE. It remains an unexplored research area with just two identified AL works (Gao et al., 2021; Rudovic et al., 2019). A generative AL sampler is proposed, MVAAL, that

successfully aligns the shared latent space of a multi-modal VAE as a PoE. This unified pipeline, inspired by AlignMM (Yang et al., 2019), leverages visual reconstructive representations from both labelled and unlabelled images. Furthermore, the aligned latent representations are combined with the learner's objective in a task-aware manner. In the end, several selection functions are evaluated in an AL setting for three multi-modal hand datasets.

7.2 LIMITATIONS AND CHALLENGES

Several limitations have been encountered during the investigation of the proposed methods. These will be revised in this section.

The first AL method that approached the 3D-HPE task, CKE, has been conducting experiments on three depth-image datasets. The Bayesian approximation approach of extracting the uncertainties has been shown to have a tedious computation. These uncertainties were generated and obtained as same-batch averages of multiple MC Dropout inferences. This intense computational effort has been reduced for the DeepPrior architecture by finding the optimal number of variational inferences. If the methodology is applied to other learners with a larger set of parameters, the computational cost will have to be mitigated by powerful hardware configurations. CKE also has an intrusive requirement of adding Dropout layers after every layer with parameters. AL developers may find themselves restricted to architectural changes of the learner.

Computational complexity is an ongoing limitation to the other proposed methods, the GCN sampler, MoBYv2AL and MVAAL. The GCN sampler with its corresponding adjacency matrix grows in the quadratic order to the number of nodes. The pool of data for selection should be acquired to fit the computational requirements. For the MoBYv2AL framework, the proposed design doubles the number of parameters with the targeted learner. On the other hand, MVAAL can be trained offline and separately from the learner. Fine-tuning can be applied with the new labelled examples for the discriminator loss. Thus, it is more optimal from a computational perspective than the other methods.

Some other met limitations of DAL from the GCN sampler have been addressed by MoBYv2AL ('cold-start' problem and the distribution shift). However, MoBYv2AL as CKE has architectural constraints and expects to have all the SSL building blocks despite the size of the learner.

Throughout the development of the proposed AL methodologies, the biggest challenge has been the deployment and the numerous runs of the

experiments. Finding the most optimal unlabelled samples is not a straightforward process. As a set of good practices, starting with some reasonable hypotheses and analysing the selections qualitatively on toy datasets has helped tremendously. If these stages were successful, scaling up to other datasets and monitoring the average performance would become easier to deploy.

7.3 FUTURE WORK

All the chapters follow a sequential research roadmap. From CKE to the GCN, SSL and VAE samplers, the active learning methodologies try to improve and expand the informativeness to the learner incrementally. Furthermore, discovered challenges on the way are also complementary overcome (like 'cold-start' or distribution shift) and kept with the parallel progress of 3D-HPE (to multi-modal pipelines). Following the limitations discussed previously, future work is still to be considered for the presented methods.

Although CKE has shown remarkable performance on the depth-based hand datasets, it will be interesting to evaluate the sampling in the multi-modal setup as in Chapter 6. Continuing with the GCN sampler, the extensive experiments have covered several applications and scenarios. In future work, the GCN architecture should address the scalability issue with a better-guided procedure rather than just a random pre-sampling.

For MoBYv2AL, the SSL-based sampler from Chapter 5 has obtained consistent performance with a large margin for the image classification experiments. Its modular query-key encoders can also be transferred to other tasks, as shown in Xie et al. (2021). As 3D-HPE is the main application in the thesis, MoBYv2AL will be adapted in future works. Thus, particular challenges will have to be investigated, like the effects of the data augmentation and how the contrastive loss converges with the memory bank of the new keys. Furthermore, the saturation nature that MoBYv2AL may have at the initial training stage will have to be investigated. For a better exploration-exploitation trade-off, likely uncertainty-based AL selection techniques combined with the contrastive visual representations may have more substantial potential.

The work presented in Chapter 6, MVAAL, is still at a concept stage and has been part of the current and future research. The multi-modal learner itself has shown to have a robust performance despite the proposed AL selections. This effect of the multi-modal sets has been shown in Gao et al. (2021) as well. The further directions should continue to analyse the learner's integration of the modalities and redefine the objective according to the

informativeness of each modality. MVAAL, as a unified AL methodology, is promising, and its hypotheses and insights will be tested on the learner again with the mentioned changes.

In terms of AL, all the pool-based scenario experiments have been run on public pre-annotated datasets. As part of future works, the proposed methodologies will be tested in real-world scenarios, especially in the situation of long-tail data distributions.

7.4 STATEMENT FROM THE AUTHOR

Active learning has been demonstrated to be an essential research direction, especially in deep learning. Optimising the data and re-enforcing learners with representative examples are just two of the key elements of active learning. Multiple domains like medical and automotive have significantly benefited from the machine or human-in-the-loop systems. This has saved time, costs, and effort during data acquisition and annotation. Moreover, it has assisted developers and researchers with various expertise to understand better the decisions of a highly parameterised learning algorithm. In the author's realisation, this work will continue to propagate ripples of insights and achievements to both deep active learning and computer vision fields.

BIBLIOGRAPHY

Abel, R. and Louzoun, Y. (2019), 'Regional based query in graph active learning'. 1906.08541v1. (pages 69, 70, 71, 78).

Agarwal, S., Arora, H., Anand, S. and Arora, C. (2020), Contextual diversity for active learning, *in* 'ECCV'. (pages 93, 100, 101).

Baek, S., Kim, K. and Kim, T.-K. (2018*a*), Augmented skeleton space transfer for depth-based hand pose estimation, *in* 'CVPR'. (page 25).

Baek, S., Kim, K. and Kim, T.-K. (2019), Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering, *in* '(CVPR)', pp. 1067–1076. (pages 44, 111, 113, 120).

Baek, S., Kim, K. I. and Kim, T.-K. (2018*b*), Augmented skeleton space transfer for depth-based hand pose estimation, *in* 'CVPR'. (pages 44, 50).

Beluch Bcai, W. H., Nürnberger, A. and Bcai, J. M. K. (2018), The power of ensembles for active learning in image classification, *in* 'CVPR'. (pages 52, 64, 69, 71, 73, 79).

Bengar, J. Z., van de Weijer, J., Twardowski, B. and Raducanu, B. (2021), Reducing label effort: Self-supervised meets active learning, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision', pp. 1631–1639. (pages 93, 95, 98, 109).

Berthelot, D., Carlini, N., Goodfellow, I., Oliver, A., Papernot, N. and Raffel, C. (2019), Mixmatch: A holistic approach to semi-supervised learning, *in* 'NeurIPS'. (pages 93, 94, 102).

Bhattarai, B., Baek, S., Bodur, R. and Kim, T.-K. (2020), Sampling strategies for gan synthetic data, *in* 'ICASSP'. (page 88).

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A. and Vandergheynst, P. (2017), 'Geometric deep learning: going beyond euclidean data', *IEEE Signal Processing Magazine* . (page 69).

Bruna, J., Zaremba, W., Szlam, A. and Lecun, Y. (2014), Spectral networks and locally connected networks on graphs, *in* 'ICLR'. (page 39).

Budd, S., Robinson, E. C. and Kainz, B. (2021), 'A survey on active learning and human-in-the-loop deep learning for medical image analysis', *Medical Image Analysis* **71**, 102062. doi: https://doi.org/10.1016/j.media.2021.102062 (pages 23, 113).

Buyu, L. and Vittorio, F. (2017), Active Learning for Human Pose Estimation, *in* 'ICCV'. Retrieved from `http://calvin.inf.ed.ac.uk/wp-content/uploads/Publications/liu17iccv.pdf` (pages 52, 56).

Cai, H., Zheng, V. W. and Chen-Chuan Chang, K. (2017), 'Active Learning for Graph Embedding'. 1705.05085v1. (pages 69, 70, 71, 78).

Cai, Y., Ge, L., Cai, J. and Yuan, J. (2018), Weakly-supervised 3d hand pose estimation from monocular rgb images, *in* 'ECCV', Springer-Verlag, p. 678–694. doi: 10.1007/978-3-030-01231-1_41 (page 113).

Caramalau, R., Bhattarai, B. and Kim, T.-K. (2021), Sequential graph convolutional network for active learning, *in* 'CVPR'. (pages 93, 100).

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P. and Joulin, A. (2020), Unsupervised learning of visual features by contrasting cluster assignments, *in* 'NeurIPS'. (page 40).

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. and Joulin, A. (2021), Emerging properties in self-supervised vision transformers, *in* 'ICCV'. (pages 93, 94, 96, 110).

Chang, J., Moon, G. and Lee, K. (2018), V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map, *in* 'CVPR'. (pages 25, 44, 45, 50, 111).

Charles, R. Q., Su, H., Kaichun, M. and Guibas, L. J. (2017), Pointnet: Deep learning on point sets for 3d classification and segmentation, *in* 'CVPR', IEEE. (page 50).

Chen, D., Lin, Y., Li, W., Li, P., Zhou, J. and Sun, X. (2020), 'Measuring and relieving the over-smoothing problem for graph neural networks from the topological view', *Proceedings of the AAAI Conference on Artificial Intelligence* **34**(04), 3438–3445. doi: 10.1609/aaai.v34i04.5747 (page 39).

Chen, T., Kornblith, S., Norouzi, M. and Hinton, G. E. (2020), A simple framework for contrastive learning of visual representations, *in* 'ICML'. (pages 40, 94, 96).

Chen, X., Fan, H., Girshick, R. and He, K. (2020), 'Improved baselines with momentum contrastive learning', *arXiv preprint arXiv:2003.04297* . (pages 93, 94, 98, 106, 110).

Chen, X. and He, K. (2021), Exploring simple siamese representation learning, *in* 'CVPR'. (pages 40, 95, 109).

Chen, X., Wang, G., Guo, H. and Zhang, C. (2020), 'Pose guided structured region ensemble network for cascaded hand pose estimation', *Neurocomputing* **395**, 138–149. doi: 10.1016/j.neucom.2018.06.097 (page 44).

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S. and Choo, J. (2018), Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, *in* 'CVPR'. (page 87).

Cohn, D. A., Atlas, L. E. and Ladner, R. E. (1994), 'Improving generalization with active learning', *Machine Learning* **15**, 201–221. (page 34).

Cohn, D. A., Ghahramani, Z. and Jordan, M. I. (1996), 'Active learning with statistical models', **4**(1), 129–145. (page 35).

Cortes, C. and Vapnik, V. (1995), 'Support-vector networks', *Machine learning* . (page 69).

Culotta, A. and McCallum, A. (2005), Reducing labeling effort for structured prediction tasks, *in* 'Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2', AAAI'05, AAAI Press, p. 746–751. (page 35).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009), Imagenet: A large-scale hierarchical image database, *in* '2009 IEEE Conference on Computer Vision and Pattern Recognition', pp. 248–255. doi: 10.1109/CVPR.2009.5206848 (page 21).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2020), 'An image is worth 16x16 words: Transformers for image recognition at scale'. (pages 97, 113).

Erol, A., Bebis, G., Nicolescu, M., Boyle, R. D. and Twombly, X. (2007), 'Vision-based hand pose estimation: A review'. (pages 23, 43, 44, 50, 53).

Fujii, A., Inui, K., Tokunaga, T. and Tanaka, H. (1998), 'Selective sampling for example-based word sense disambiguation', *Computational Linguist-*

ics **24**(4), 573–597. Retrieved from https://aclanthology.org/J98-4002 (page 36).

Fukushima, K. (2004), 'Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position', *Biological Cybernetics* **36**, 193–202. (page 21).

Gal, Y. and Ghahramani, Z. (2015), 'Bayesian convolutional neural networks with bernoulli approximate variational inference'. (pages 37, 51).

Gal, Y. and Ghahramani, Z. (2016), Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, *in* 'ICML'. (pages 51, 52, 54, 100, 106).

Gal, Y., Islam, R. and Ghahramani, Z. (2017), Deep Bayesian Active Learning with Image Data, *in* 'ICML'. (pages 52, 56, 69, 71).

Gao, L., Yang, H., Zhou, C., Wu, J., Pan, S. and Hu, Y. (2018), Active discriminative network representation learning, *in* 'IJCAI'. (pages 69, 70, 71, 78).

Gao, L., Zhou, R., Dong, C., Feng, C., Li, Z., Wan, X. and Liu, L. (2021), Multi-modal active learning for automatic liver fibrosis diagnosis based on ultrasound shear wave elastography, *in* '2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)', pp. 410–414. doi: 10.1109/ISBI48211.2021.9434170 (pages 114, 122, 126, 128).

Gao, M., Zhang, Z., Yu, G., Arık, S., Davis, L. and Pfister, T. (2020), Consistency-based semi-supervised active learning: Towards minimizing labeling cost, *in* 'ECCV'. (pages 35, 93, 95, 102).

Garcia-Hernando, G., Yuan, S., Baek, S. and Kim, T.-K. (2018), First-person hand action benchmark with rgb-d videos and 3d hand pose annotations, *in* 'CVPR'. (page 26).

Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., Cai, J. and Yuan, J. (2019), 3d hand shape and pose estimation from a single rgb image, *in* 'CVPR', pp. 10825–10834. (pages 44, 111, 113, 120).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014), Generative adversarial nets, *in* 'NeurIPS'. (page 87).

Goodfellow, I. J., Bulatov, Y., Ibarz, J., Arnoud, S. and Shet, V. (2013*a*), 'Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks'. 1312.6082v4. (pages 22, 100).

Goodfellow, I. J., Bulatov, Y., Ibarz, J., Arnoud, S. and Shet, V. (2013*b*), 'Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks'. 1312.6082v4. (page 78).

Gorriz, M., Carlier, A., Faure, E. and Giró-i-Nieto, X. (2017), 'Cost-effective active learning for melanoma segmentation', *CoRR* **abs/1711.09168**. (pages 100, 106, 114).

Goyal, P., Caron, M., Lefaudeux, B., Xu, M., Wang, P., Pai, V., Singh, M., Liptchinsky, V., Misra, I., Joulin, A. and Bojanowski, P. (2021), 'Self-supervised pretraining of visual features in the wild'. https://arxiv.org/abs/2103.01988 (page 41).

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R. and Valko, M. (2020), Bootstrap your own latent - a new approach to self-supervised learning, *in* 'NeurIPS'. (pages 40, 93, 94, 96, 97, 106, 110).

Guo, H., Wang, G., Chen, X. and Zhang, C. (2017), 'Towards good practices for deep 3d hand pose estimation'. https://arxiv.org/abs/1707.07248 (page 44).

Hamilton, W. L., Ying, R. and Leskovec, J. (2017), Inductive representation learning on large graphs, *in* 'NeurIPS'. (page 90).

Har-Peled, S. and Kushal, A. (2005), Smaller coresets for k-median and k-means clustering, *in* 'SCG'. (page 71).

Haussmann, E., Fenzi, M., Chitta, K., Ivanecky, J., Xu, H., Roy, D., Mittel, A., Koumchatzky, N., Farabet, C. and Alvarez, J. M. (2020), 'Scalable active learning for object detection', *CoRR* **abs/2004.04699**. Retrieved from https://arxiv.org/abs/2004.04699 (pages 23, 113).

He, K., Fan, H., Wu, Y., Xie, S. and Girshick, R. B. (2020), Momentum contrast for unsupervised visual representation learning, *in* 'CVPR'. (pages 40, 93, 94, 96, 97, 98).

He, K., Zhang, X., Ren, S. and Sun, J. (2016), Deep residual learning for image recognition, *in* 'CVPR'. (pages 44, 73, 77, 100, 113, 115).

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S. and Lerchner, A. (2017), beta-VAE: Learning basic visual concepts with a constrained variational framework, *in* 'ICLR'. Retrieved from https://openreview.net/forum?id=Sy2fzU9gl (pages 41, 117).

Hinton, G. E., Osindero, S. and Teh, Y.-W. (2006), 'A fast learning algorithm for deep belief nets', *Neural Comput.* **18**(7), 1527–1554. doi: 10.1162/neco.2006.18.7.1527 (page 21).

Houlsby, N., Huszár, F., Ghahramani, Z. and Lengyel, M. (2011), 'Bayesian Active Learning for Classification and Preference Learning'. 1112.5745v1. (page 71).

Hsiao, D.-Y., Sun, M., Ballweber, C., Cooper, S. and Popović, Z. (2016), Proactive sensing for improving hand pose estimation, *in* 'Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems'. doi: 10.1145/2858036.2858587 (pages 24, 52).

Huang, F., Zeng, A., Liu, M., Qin, J. and Xu, Q. (2018), Structure-aware 3d hourglass network for hand pose estimation from single depth image, *in* 'BMVC'. (page 44).

Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A. and Carreira, J. (2021), Perceiver: General perception with iterative attention, *in* M. Meila and T. Zhang, eds, 'ICML', Vol. 139 of *Proceedings of Machine Learning Research*, PMLR, pp. 4651–4664. Retrieved from https://proceedings.mlr.press/v139/jaegle21a.html (page 111).

Jesionkowska, J., Wild, F. and Deval, Y. (2020), 'Active learning augmented reality for steam education—a case study', *Education Sciences* **10**(8). doi: 10.3390/educsci10080198 (page 23).

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. K. (1999), *An Introduction to Variational Methods for Graphical Models*, MIT Press, Cambridge, MA, USA, p. 105–161. (page 37).

Joulin, A., Grave, E., Bojanowski, P. and Mikolov, T. (2016), 'Bag of tricks for efficient text classification'. https://arxiv.org/abs/1607.01759 (page 40).

Kato, H. and Takemura, K. (2016), Hand pose estimation based on active bone-conducted sound sensing, *in* 'Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Ad-

junct', UbiComp '16, Association for Computing Machinery, p. 109–112. doi: 10.1145/2968219.2971403 (page 24).

Kendall, A., Badrinarayanan, V. and Cipolla, R. (2015), 'Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding'. (pages 53, 55, 60).

Kendall, A. and Gal, Y. (2017), What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, *in* 'NeurIPS'. (pages 26, 37, 49, 51, 52, 53, 55, 56, 61).

Kendall, A., Gal, Y. and Cipolla, R. (2018), Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, *in* 'CVPR'. (page 51).

Kim, K., Park, D., Kim, K. and Chun, S. (2021), Task-aware variational adversarial active learning, *in* 'CVPR', IEEE Computer Society, pp. 8162–8171. doi: 10.1109/CVPR46437.2021.00807 (pages 35, 93).

Kingma, D. P. and Lei Ba, J. (2015), ADAM: A Method for Stochastic Optimization, *in* 'ICLR'. (pages 60, 77, 88).

Kipf, T. N. and Welling, M. (2017), Semi-supervised classification with graph convolutional networks, *in* 'ICLR'. (pages 39, 69, 70, 75).

Kirsch, A., Rainforth, T. and Gal, Y. (2021), Test distribution-aware active learning: A principled approach against distribution shift and outliers. (page 106).

Kirsch, A., Van Amersfoort, J. and Gal, Y. (2019), BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning, *in* 'NeurIPS'. (page 71).

Kontorovich, A., Sabato, S. and Urner, R. (2016), Active nearest-neighbor learning in metric spaces, *in* 'NeurIPS'. (page 71).

Krizhevsky, A. (2012), 'Learning multiple layers of features from tiny images', *University of Toronto* . (pages 22, 78, 100).

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012), Imagenet classification with deep convolutional neural networks, *in* 'NeurIPS'. (pages 21, 25).

Lang, K. and Baum, E. (1992), Query learning can work poorly when a human oracle is used, *in* 'Proceedings of the IEEE International Joint Conference on Neural Networks'. (page 34).

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T. and Van Knippenberg, A. (2010), 'Presentation and validation of the radboud faces database', *Cognition and emotion* . (page 87).

LeCun, Y. and Cortes, C. (2010), 'MNIST handwritten digit database'. (pages 22, 52).

Lewis, D. D. and Gale, W. A. (1994), A sequential algorithm for training text classifiers, *in* 'SIGIR', p. 3–12. (pages 34, 35).

Li, J., Xiong, C. and Hoi, S. C. (2021), Semi-supervised learning with contrastive graph regularization, *in* 'ICCV'. (page 94).

Luo, J., Wang, J., Cheng, N. and Xiao, J. (2021), Loss prediction: End-to-end active learning approach for speech recognition, *in* '2021 International Joint Conference on Neural Networks (IJCNN)', pp. 1–7. doi: 10.1109/IJCNN52387.2021.9533839 (page 23).

Maltz, D. and Ehrlich, K. (1995), Pointing the way: Active collaborative filtering., pp. 202–209. doi: 10.1145/223904.223930 (page 27).

Margatina, K., Vernikos, G., Barrault, L. and Aletras, N. (2021), Active learning by acquiring contrastive examples, *in* 'EMNLP'. (page 95).

McCallum, A. and Nigam, K. (1998), Employing em and pool-based active learning for text classification, *in* 'ICML', ICML '98, Morgan Kaufmann Publishers Inc., p. 350–358. (page 36).

Melax, S., Keselman, L. and Orsten, S. (2013), Dynamics based 3d skeletal hand tracking, *in* 'Proceedings of Graphics Interface 2013', GI '13, Canadian Information Processing Society, p. 63–70. (page 47).

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013), 'Efficient estimation of word representations in vector space'. https://arxiv.org/abs/1301.3781 (page 40).

Moon, G., Yu, S.-i., Wen, H., Shiratori, T. and Lee, K. M. (2020), Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image, *in* 'ECCV'. (pages 24, 26, 112).

Nair, V. and Hinton, G. E. (2010), Rectified linear units improve restricted boltzmann machines, *in* 'ICML'. (page 75).

Newell, A., Yang, K. and Deng, J. (2016), Stacked hourglass networks for human pose estimation, *in* 'ECCV'. (page 46).

Nguyen, H. T. and Smeulders, A. W. M. (2004), 'Active learning using pre-clustering', *ICML* . (page 36).

Nie, X., Zhang, J., Yan, S. and Feng, J. (2019), Single-stage multi-person pose machines, *in* 'ICCV'. (page 24).

Nix, D. and Weigend, A. (1994), Estimating the mean and variance of the target probability distribution, *in* 'Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)', Vol. 1, pp. 55–60 vol.1. doi: 10.1109/ICNN.1994.374138 (page 38).

Northcutt, C. G., Athalye, A. and Mueller, J. (2021), Pervasive label errors in test sets destabilize machine learning benchmarks, *in* 'Proceedings of the 35th Conference on Neural Information Processing Systems Track on Datasets and Benchmarks'. (page 22).

Oberweger, M. and Lepetit, V. (2017), Deepprior++: Improving fast and accurate 3d hand pose estimation, *in* 'ICCV'. (pages 25, 44, 46, 50, 111, 114, 115).

Oberweger, M., Wohlhart, P. and Lepetit, V. (2015), Hands deep in deep learning for hand pose estimation, *in* 'CVWW'. (pages 25, 44, 45, 50, 51, 54, 60, 73, 88).

Oord, A. v. d., Li, Y. and Vinyals, O. (2018), 'Representation learning with contrastive predictive coding', *arXiv.org perpetual, non-exclusive license* . (pages 40, 97).

Osband, I. (2016), Risk versus uncertainty in deep learning : Bayes , bootstrap and the dangers of dropout, *in* 'NeurIPS Workshop'. (pages 52, 55).

Page, L., Brin, S., Motwani, R. and Winograd, T. (1999), 'The pagerank citation ranking: Bringing order to the web.'. (page 71).

Pennington, J., Socher, R. and Manning, C. (2014), Glove: Global vectors for word representation, *in* 'EMNLP', Vol. 14, pp. 1532–1543. doi: 10.3115/v1/D14-1162 (page 40).

Pinsler, R., Gordon, J., Nalisnick, E. and Miguel Hernandez-Lobato, J. (2019), Bayesian Batch Active Learning as Sparse Subset Approximation, *in* 'NeurIPS'. (pages 56, 71).

Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A. and Carin, L. (2016), Variational autoencoder for deep learning of images, labels and captions, *in* 'NeurIPS'. (page 70).

Rad, M., Oberweger, M. and Lepetit, V. (2018), Feature mapping for learning fast and accurate 3d pose inference from synthetic images, *in* 'CVPR', pp. 4663–4672. (page 44).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I. (2021), Learning transferable visual models from natural language supervision, *in* M. Meila and T. Zhang, eds, 'ICML', Vol. 139 of *Proceedings of Machine Learning Research*, PMLR, pp. 8748–8763. Retrieved from https://proceedings.mlr.press/v139/radford21a.html (page 111).

Ravuri, S. and Vinyals, O. (2019), Seeing is not necessarily believing: Limitations of biggans for data augmentation, *in* 'ICLR'. (page 88).

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X. and Wang, X. (2020), 'A survey of deep active learning'. https://arxiv.org/abs/2009.00236 (page 36).

Romero, J., Tzionas, D. and Black, M. J. (2017), 'Embodied hands', *ACM Transactions on Graphics* **36**(6), 1–17. doi: 10.1145/3130800.3130883 (pages 48, 113).

Ronneberger, O., Fischer, P. and Brox, T. (2015), 'U-net: Convolutional networks for biomedical image segmentation', *MICCAI* . (pages 59, 90).

Rowlinson, P. (1996), 'D. m. cvetković, m. doob and h. sachs spectra of graphs (3rd edition) (johann ambrosius barth verlag, heidelberg-leipzig 1995), 447 pp., 3 335 00407 8, dm 168.', *Proceedings of the Edinburgh Mathematical Society* **39**(1), 188–189. doi: 10.1017/S0013091500022902 (page 39).

Rudovic, O., Zhang, M., Schuller, B. and Picard, R. (2019), Multi-modal active learning from human data: A deep reinforcement learning approach, *in* '2019 International Conference on Multimodal Interaction', ICMI '19, Association for Computing Machinery, p. 6–15. doi: 10.1145/3340555.3353742 (pages 114, 126).

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986), 'Learning representations by back-propagating errors', *Nature* **323**, 533–536. (page 21).

Sanchez-Riera, J., Srinivasan, K., Hua, K.-L., Cheng, W.-H., Hossain, M. A. and Alhamid, M. F. (2018), 'Robust rgb-d hand tracking using deep learning priors', *IEEE Transactions on Circuits and Systems for Video Technology* **28**(9), 2289–2301. doi: 10.1109/TCSVT.2017.2718622 (page 44).

Scheffer, T., Decomain, C. and Wrobel, S. (2001), Active hidden markov models for information extraction, *in* 'IDA'. (page 35).

Schein, A. I., Popescul, A., Ungar, L. H. and Pennock, D. M. (2002), Methods and metrics for cold-start recommendations, *in* 'SIGIR '02', Association for Computing Machinery, p. 253–260. doi: 10.1145/564376.564421 (page 27).

Schumann, R. and Rehbein, I. (2019), Active learning via membership query synthesis for semi-supervised sentence classification, *in* 'Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)', Association for Computational Linguistics, pp. 472–481. doi: 10.18653/v1/K19-1044 (page 34).

Sener, O. and Savarese, S. (2018), Active Learning for Convolutional Neural Networks: A Core-set approach, *in* 'ICLR'. arXiv: 1708.00489v4 (pages 13, 50, 52, 56, 57, 58, 61, 70, 71, 72, 73, 76, 78, 85, 88, 100).

Settles, B. (2009), Active learning literature survey, Computer Sciences Technical Report 1648, University of Wisconsin–Madison. (pages 34, 56, 72).

Settles, B. and Craven, M. (2008), An analysis of active learning strategies for sequence labeling tasks, *in* 'Proceedings of the Conference on Empirical Methods in Natural Language Processing', EMNLP '08, Association for Computational Linguistics, p. 1070–1079. (page 35).

Seung, H. S., Opper, M. and Sompolinsky, H. (1992), Query by committee, *in* 'Proceedings of the Fifth Annual Workshop on Computational Learning Theory', COLT '92, Association for Computing Machinery, p. 287–294. doi: 10.1145/130385.130417 (page 35).

Shannon, C. E. (1948), 'A mathematical theory of communication', *The Bell System Technical Journal* . (pages 35, 114).

Sharp, T., Wei, Y., Freedman, D., Kohli, P., Krupka, E., Fitzgibbon, A., Izadi, S., Keskin, C., Robertson, D., Taylor, J., Shotton, J., Kim, D., Rhemann, C., Leichter, I. and Vinnikov, A. (2015), Accurate, robust, and flexible real-time hand tracking. (pages 26, 45).

Simon, T., Joo, H. and Sheikh, Y. (2017), 'Hand keypoint detection in single images using multiview bootstrapping', *CVPR* . (page 44).

Simonyan, K. and Zisserman, A. (2015), Very Deep Convolutional Network for Large-scale image recognition, *in* 'ICLR'. (pages 25, 73, 85, 100, 150).

Sinha, S., Ebrahimi, S. and Darrell, T. (2019), Variational Adversarial Active Learning, *in* 'ICCV'. (pages 30, 35, 52, 56, 61, 69, 70, 73, 74, 78, 79, 88, 93, 100, 101).

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, *in* 'NeurIPS2020'. (page 94).

Spurr, A., Dahiya, A., Wang, X., Zhang, X. and Hilliges, O. (2021), Self-supervised 3d hand pose estimation from monocular rgb via contrastive learning, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision', pp. 11230–11239. (page 110).

Spurr, A., Song, J., Park, S. and Hilliges, O. (2018), Cross-modal deep variational hand pose estimation, *in* 'CVPR', pp. 89–98. (pages 41, 111, 113, 117).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014), 'Dropout: A simple way to prevent neural networks from overfitting', *Journal of Machine Learning Research* **15**(56), 1929–1958. Retrieved from `http://jmlr.org/papers/v15/srivastava14a.html` (pages 51, 54).

Supancic, J. S., Rogez, G., Yang, Y., Shotton, J. and Ramanan, D. (2015), Depth-based hand pose estimation: Data, methods, and challenges, *in* 'ICCV'. (page 26).

Tang, D., Jin Chang, H., Tejani, A. and Kim, T.-K. (2014), Latent regression forest: Structured estimation of 3d articulated hand posture, *in* 'CVPR'. (pages 14, 25, 26, 29, 44, 45, 46, 50, 59, 74, 90, 111).

Tang, D., Taylor, J., Kohli, P., Keskin, C., Kim, T.-K. and Shotton, J. (2015), Opening the black box: Hierarchical sampling optimization for estimating human hand pose, *in* 'ICCV', pp. 3325–3333. doi: 10.1109/ICCV.2015.380 (page 44).

Tang, D., Yu, T. and Kim, T. (2013), Real-time articulated hand pose estimation using semi-supervised transductive regression forests, *in* 'ICCV'. (page 25).

Tarvainen, A. and Valpola, H. (2017), Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, *in* 'NeurIPS'. (page 94).

Tompson, J., Stein, M., Lecun, Y. and Perlin, K. (2014), 'Real-time continuous pose recovery of human hands using convolutional networks', *ACM Transactions on Graphics* . (pages 26, 29, 45, 46, 50, 56, 59, 90, 111).

Tsang, I. W., Kwok, J. T. and Cheung, P.-M. (2005), 'Core vector machines: Fast svm training on very large data sets', *JMLR*. . (page 71).

van der Maaten, L. and Hinton, G. (2008), 'Visualizing data using t-sne'. JMLR. (pages 82, 108).

Wan, C., Probst, T., Van Gool, L. and Yao, A. (2018), Dense 3d regression for hand pose estimation, *in* 'CVPR'. (page 50).

Wolf, G. (2011), Facility location: concepts, models, algorithms and case studies, *in* 'Contributions to Management Science'. (pages 57, 76, 100).

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K. and Vajda, P. (2020), 'Visual transformers: Token-based image representation and processing for computer vision'. (pages 97, 113).

Wu, M. and Goodman, N. (2018), Multimodal generative models for scalable weakly-supervised learning, *in* 'NeurIPS', NIPS'18, p. 5580–5590. (page 117).

Wu, Y., Xu, Y., Singh, A., Yang, Y. and Dubrawski, A. (2019), 'Active Learning for Graph Neural Networks via Node Feature Propagation'. 1910.07567v1. (pages 69, 71, 76, 78, 82, 88).

Wu, Z., Xiong, Y., Stella, X. Y. and Lin, D. (2018), Unsupervised feature learning via non-parametric instance discrimination, *in* 'CVPR'. (page 98).

Xiao, H., Rasul, K. and Vollgraf, R. (2017), 'Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms'. 1708.07747v2. (pages 78, 100).

Xie, Z., Lin, Y., Yao, Z., Zhang, Z., Dai, Q., Cao, Y. and Hu, H. (2021), 'Self-supervised learning with swin transformers', *arXiv preprint arXiv:2105.04553* . (pages 30, 40, 41, 93, 94, 96, 97, 99, 106, 128).

Xiong, F., Zhang, B., Xiao, Y., Cao, Z., Yu, T., Zhou, J. T. and Yuan, J. (2019), A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image, *in* 'ICCV'. (pages 44, 50, 111).

Yang, L., Li, S., Lee, D. and Yao, A. (2019), Aligning latent spaces for 3d hand pose estimation, *in* 'ICCV', pp. 2335–2343. doi: 10.1109/ICCV.2019.00242 (pages 31, 41, 44, 111, 112, 113, 115, 117, 127).

Ye, Q., Yuan, S. and Kim, T.-k. (2016), Spatial Attention Deep Net with Partial PSO for Hierarchical Hybrid Hand Pose Estimation, *in* 'ECCV'. (pages 25, 44, 45, 50, 111, 115).

Yoo, D. and Kweon, I. S. (2019), Learning Loss for Active Learning, *in* 'CVPR'. (pages 35, 52, 56, 64, 69, 70, 73, 78, 79, 88, 93, 100).

Yuan, S., Garcia-Hernando, G., Stenger, B., Moon, G., Chang, J. Y., Lee, K. M., Molchanov, P., Kautz, J., Honari, S., Ge, L., Yuan, J., Chen, X., Wang, G., Yang, F., Akiyama, K., Wu, Y., Wan, Q., Madadi, M., Escalera, S., Li, S., Lee, D., Oikonomidis, I., Argyros, A. and Kim, T.-K. (2018), Depth-based 3d hand pose estimation: From current achievements to future goals, *in* 'CVPR'. (page 26).

Yuan, S., Ye, Q., Garcia-Hernando, G. and Kim, T.-K. (2017*a*), 'The 2017 hands in the million challenge on 3d hand pose estimation'. (pages 26, 45).

Yuan, S., Ye, Q., Garcia-Hernando, G. and Kim, T.-K. (2017*b*), The 2017 hands in the million challenge on 3d hand pose estimation, *in* 'ICCV'. (pages 47, 111).

Yuan, S., Ye, Q., Stenger, B., Jain, S. and Kim, T.-K. (2017), Bighand2.2m benchmark: Hand pose dataset and state of the art analysis, *in* 'CVPR'. (pages 26, 29, 46, 47, 49, 50, 59, 90, 111, 112).

Zagoruyko, S. and Komodakis, N. (2016), Wide residual networks, *in* 'BMVC'. (pages 102, 113).

Zeng, H., Zhou, H., Srivastava, A., Kannan, R. and Prasanna, V. K. (2020), Graphsaint: Graph sampling based inductive learning method, *in* 'ICLR'. (page 90).

Zeng, J., Lesnikowski, A. and Alvarez, J. M. (2018), 'The relevance of bayesian layer positioning to model uncertainty in deep bayesian active learning'. (page 51).

Zhang, J., Jiao, J., Chen, M., Qu, L., Xu, X. and Yang, Q. (2017), A hand pose tracking benchmark from stereo matching, *in* 'ICIP', pp. 982–986. doi: 10.1109/ICIP.2017.8296428 (pages 31, 47, 112).

Zhao, L. and Akoglu, L. (2020), 'Pairnorm: Tackling oversmoothing in gnns', *ICLR* . (pages 39, 77).

Zimmermann, C. and Brox, T. (2017), Learning to estimate 3d hand pose from single rgb images, Technical report, arXiv:1705.01389. https://arxiv.org/abs/1705.01389. Retrieved from `https://lmb.informatik.uni-freiburg.de/projects/hand3d/` (pages 31, 44, 47, 112).

Zimmermann, C., Ceylan, D., Yang, J., Russel, B., Argus, M. and Brox, T. (2019), Freihand: A dataset for markerless capture of hand pose and shape from single rgb images, *in* 'ICCV'. Retrieved from "https://lmb.informatik.uni-freiburg.de/projects/freihand/" (pages 31, 47, 112).

A

ACTIVE LEARNING QUALITATIVE ANALYSIS FOR 3D HAND POSE ESTIMATION

This Appendix aims to display and briefly discuss some of the selection analysis conducted for the CKE algorithm from Chapter 3. The analysis is conducted on a version of the BigHand2.2M dataset where for every ten samples, one is selected. This results in a dataset of 220,000 images. Even though the dataset is drastically reduced, the viewpoint and articulation space is still diverse enough.

The first experiment, represented in Figure A.1, runs a simulated experiment where the unlabelled pool of data has samples uniformly distributed from camera angles. The initial labelled set is also uniform, having the same number of images from every angle.

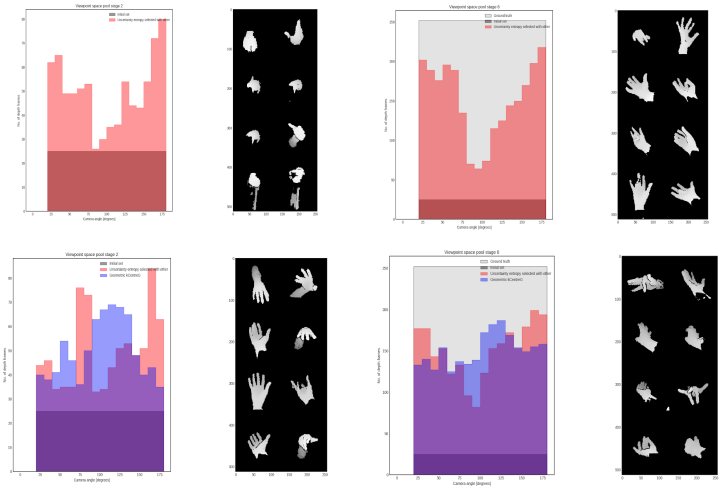

Figure A.1: Qualitative AL selection. Comparison between entropy uncertainty and CoreSet according to viewpoint and articulation distributions on BigHand2.2M

The distribution of selected images is analysed at 20% and 60% annotated data. The images shown are the top selected poses by either uncertainty sampling (top) or CoreSet (bottom). It can be clearly denoted that uncertainty sampling on the Bayesian approximation selects samples from difficult viewpoints, as it is shown on the side peaks. This behaviour is kept consistent to the last AL cycle. As the model converges to hard out-of-distribution samples, at later stages, the new sets show easy front poses (Figure A.1 top right).

On the bottom side, CoreSet, displayed in blue, chooses more frontal samples than the uncertainty estimation. This confirms one of the discussed hypotheses, where the learner converges quicker at first AL cycles. Gradually, the method starts to pick more complex examples, showing a camera angle distribution closer to uniform.

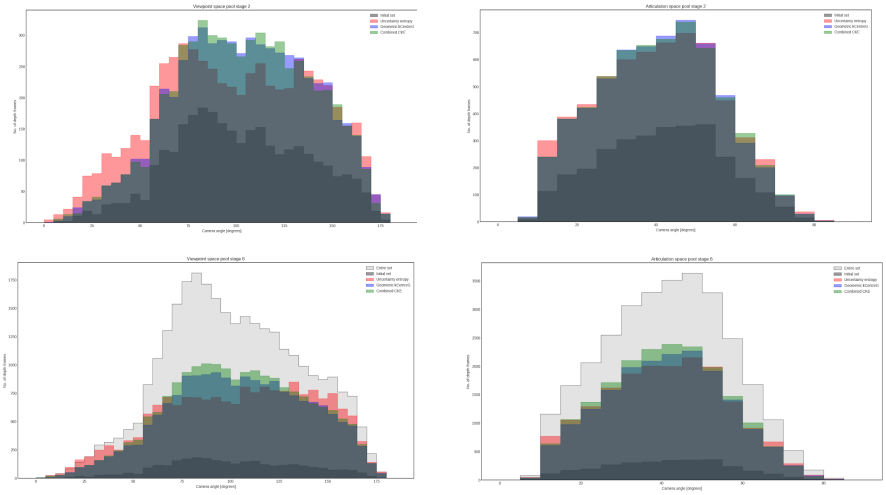

Figure A.2: Qualitative AL selection. CKE selected samples according to viewpoint and articulation distributions on BigHand2.2M

The second experiment brings the proposed method, CKE, into perspective in the distributions from Figure A.2. As in the previous experiment, the other sampling techniques are used, and the distributions are shown at 20% and 60% annotations. Distinctively from the previous experiment, articulation distributions are added to the viewpoint ones. To use the highly diverse BigHand2.2M, the entire sub-sampled set is used as an unlabelled pool. The initial training set is randomly selected (distribution displayed with grey).

In the four representations, the distributions overlap, and similar behaviours as in Figure A.1 are observed. Thus, the uncertainty sampling (red) targets outliers from the first selection stage, while CoreSet (blue) follows the initial unlabelled set distribution. The proposed CKE method (green) demonstrates an expected behaviour. At 20%, it has a close trend to CoreSet, but at 60%, the epistemic uncertainty pushes to select samples at intersections with uncertainty sampling. In this analysis, CKE adapts dynamically to expand the exploration of CoreSet.

B

EXTENDED EXPERIMENTS FOR THE GCN-BASED SAMPLER

This Appendix expands the list of experiments from Chapter 4. As in MoBYv2AL (Chapter 5), it simulates a CIFAR-10 imbalanced dataset selection, followed by a set of ablation studies and some qualitative representations.

B.1 CIFAR-10 IMBALANCED DATASET

Although, before selection, the unlabelled samples are randomised to a subset, the dataset is still relatively balanced for each class distribution. However, this is only sometimes the case where there is no prior information related to the data space. Therefore, an imbalanced CIFAR-10 is simulated in a quantitative experiment.

Beforehand the 50,000 training set was considered unlabeled, given 5,000 samples for each of the ten categories. The dataset is custom so that 5 of the ten classes contain 10 % of their original data (500 samples each). Therefore, the new unlabelled pool is composed of 27,500 images. The experiment architecture and settings are similar to the one on the full scale.

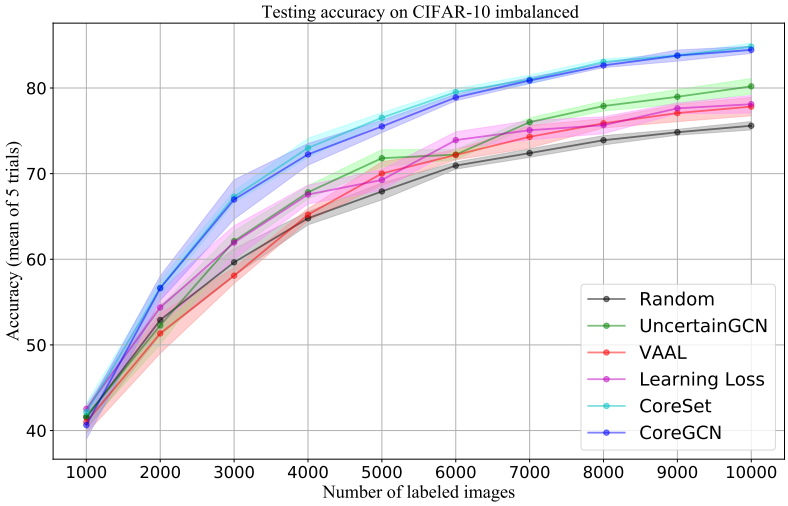

Figure B.1: Quantitative results - CIFAR-10 imbalanced dataset

Figure B.1 shows the progressions of the presented baselines. The proposed methods, UncertainGCN and CoreGCN, out-stand once again the other model-based selections like VAAL and Learning Loss. UncertainGCN scores

2% more than those methods with 80.05% mean average accuracy at 10,000 labelled samples. Meanwhile, CoreGCN achieves 84.5% top performance together with CoreSet. Thus, the geometric information is more useful in scenarios where the dataset is imbalanced.

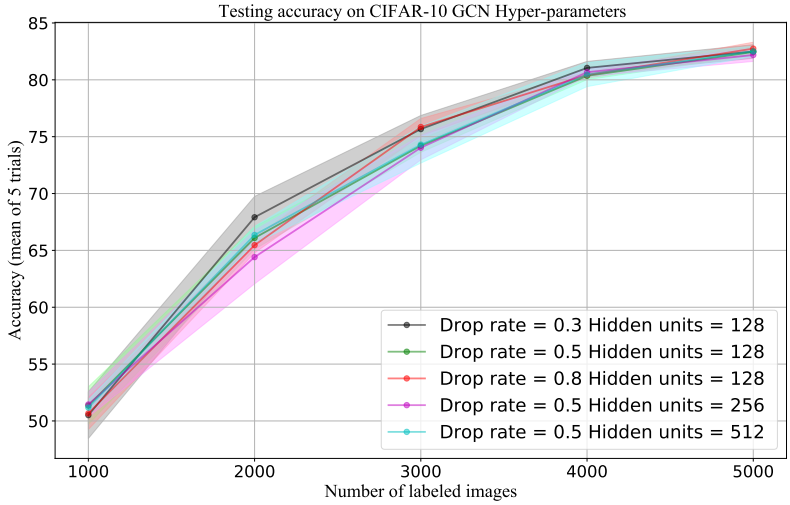

Figure B.2: Ablation studies - CIFAR-10 GCN Hyper-parameters tuning

B.2 Ablation Study - GCN Parameter Search

While varying the architectural parameters of the GCN binary classifier, a poorer selection was encountered with the Dropout rate increase from 0.3 to 0.5 or 0.8. However, when changing the size of the hidden units to 256 and 512, the UncertainGCN sampling was not affected on CIFAR-10. This might require further optimisation for different datasets, although robustness is being shown.

B.3 VGG-11 Learner for CIFAR-10 Image Classification for 3 Selection Stages

In Figure B.3, the architecture of the learner is modified from the CIFAR-10 experiment to VGG-11 (Simonyan and Zisserman, 2015). In training the VGG-11 network, the same hyper-parameters were kept. The proposed methods present robustness to this change whilst the GCN settings were left unchanged. Hence, all state-of-the-art were surpassed at this early stage. This also demonstrates how the batch size and the feature representation play an

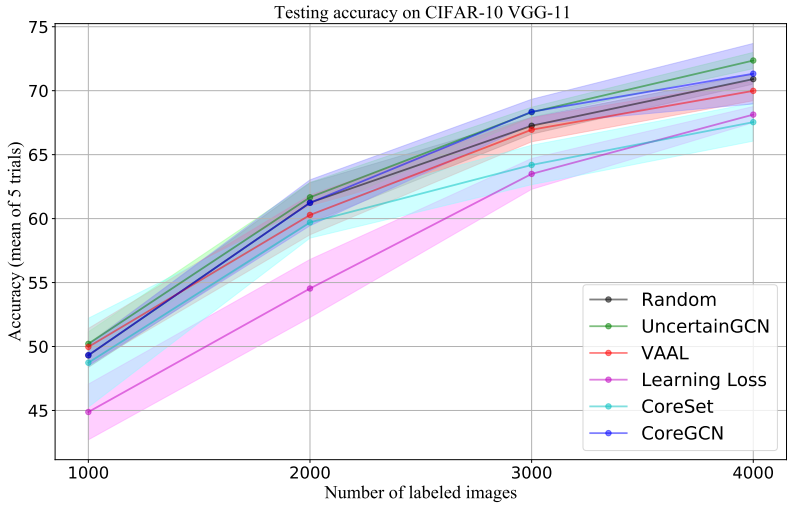

Figure B.3: CIFAR-10 Learner VGG-11 - 3 selection stages

essential role in the performances of the other baselines. The most affected baseline in this context is CoreSet.

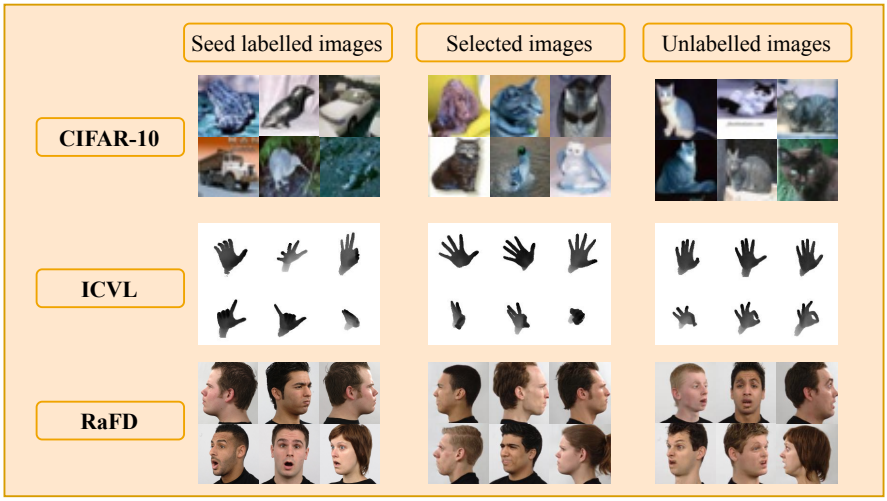

Figure B.4: Extended qualitative analysis on labelled/unlabelled images at the last selection stage for CIFAR-10, ICVL and RaFD

B.4 EXTENDED QUALITATIVE ANALYSIS ON THE AL METHOD

In Figure B.4, the qualitative analysis is extended by visualising the initial, the unlabelled and the last selected samples from CIFAR-10, ICVL and RaFD. The last selection stage for CIFAR-10 and ICVL is the 10th, while the synthetic RaFD experiment is the 4th. The seed-labelled images are acquired randomly

before the first selection stage. The RaFD seed examples are from the entire training set as the AL selection is applied on StarGAN-generated images. For all three benchmarks, the selected examples were evaluated with the proposed AL method, UncertainGCN. Although the seed-labelled samples for CIFAR-10 are randomly selected, the top query images from the "cat" class consist of difficult examples. On the other hand, the remained unlabelled images present distinguishable features, easy for the learner to predict. However, in the ICVL dataset case, the selected samples show closer and easier hand articulations compared to the initial labelled set. This is because of the highly complex set that was used as seed examples. The unlabelled images might have a lack of representativeness in the learner's perception after all the 10 sampling stages. Finally, in the RaFD synthetic sub-sampling process, it can be clearly denoted that the noisy images were left unlabelled. These present more artefacts than the selected group.

C

PUBLICATIONS OF THE AUTHOR

The following appendix lists the references and the key ideas from all the works published in the course of the PhD. These publications are also the founding pillars of the content of this thesis:

Caramalau, R., Bhattarai, B., and Kim, T.-K. (2021), Active Learning for Bayesian 3D Hand Pose Estimation, *in* 'Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)', pp. 3419–3428, .

- Introduces a Bayesian approximation for 3D-HPE to model uncertainties on regression-based learners. The Bayesian framework is designed as part of the active learning selection process. A novel sampling mechanism is proposed that combines both data representations from the learner and its corresponding model-dependent uncertainty. Moreover, a data-dependent uncertainty further helps the pose estimator to balance out difficult or noisy examples. The contributions are tested and concluded on three depth hand datasets by providing the lowest joint errors with less data.

- Chapter 3 is based on this paper.

Caramalau, R., Bhattarai, B., and Kim, T.-K. (2021), Sequential Graph Convolutional Network for Active Learning, *in* 'Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)', pp. 9583–9592, .

- Presents a novel pool-based AL method that incorporates a sequential GCN. The advantage of this method is that it obtains higher-order representation through message passing between labelled and unlabelled examples. This helps in increasing the informativeness during the selection stage either by classic uncertainty sampling or by geometric data distributions. The method is task-aware and modular, allowing it to be easily deployed to various applications.

- Chapter 4 is based on this paper.

Caramalau, R., Bhattarai, B., Stoyanov, D., and Kim, T.-K. (2022), MoBYv2AL: Self-supervised Active Learning for Image Classification, *to be presented at* 'The 33rd British Machine Vision Conference (BMVC)'.

- Introduces an AL method that addresses prior known issues in the pool-based selection: the cold-start problem and distribution shift. This is being overcome by leveraging the visual concepts of the unlabelled examples in an end-to-end self-supervised training pipeline. The contribution lies in the AL framework MoBYv2AL that promotes contrastive learning with end-task classification objectives. It makes use of prior concepts of two popular contrastive learning techniques, MoCo and BYOL. Furthermore, the MoBYv2AL deploys CoreSet to benefit the enriched representation for more informative subsets. The method achieves state-of-the-art to date on four classification datasets.
- Chapter 5 is based on this paper.