REPUBLIC OF TÜRKİYE
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

VIDEO CLASSIFICATION AND RETRIEVAL WITH LOW - DATA REGIME LEARNING

Erol ÇITAK

DOCTOR OF PHILOSOPHY THESIS
Department of Computer Engineering
Program of Computer Engineering

Supervisor
Prof. Dr. Mine Elif KARSLIGİL

August, 2024

REPUBLIC OF TÜRKİYE
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

VIDEO CLASSIFICATION AND RETRIEVAL WITH LOW-DATA REGIME LEARNING

A thesis submitted by Erol ÇITAK in partial fulfillment of the requirements for the degree of **DOCTOR OF PHILOSOPHY** is approved by the committee on 15.08.2024 in Department of Computer Engineering, Program of Computer Engineering.

Prof. Dr. Mine Elif KARSLIGİL
Yıldız Technical University
Supervisor

Approved By the Examining Committee

Prof. Dr. Mine Elif KARSLIGİL, Supervisor
Yıldız Technical University _____

Prof. Dr. Gökhan BİLGİN, Member
Yıldız Technical University _____

Prof. Dr. Hazım Kemal EKENEL, Member
İstanbul Technical University _____

Prof. Dr. Songül VARLI, Member
Yıldız Technical University _____

Prof. Dr. Berrin YANIKOĞLU, Member
Sabancı University _____

I hereby declare that I have obtained the required legal permissions during data collection and exploitation procedures, that I have made the in-text citations and cited the references properly, that I haven't falsified and/or fabricated research data and results of the study and that I have abided by the principles of the scientific research and ethics during my Thesis Study under the title of Video Classification And Retrieval With Low Data Regime Learning supervised by my supervisor, Prof. Dr. Mine Elif KARSLIGİL. In the case of a discovery of false statement, I am to acknowledge any legal consequence.

Erol ÇITAK

Signature

Dedicated to my wife and family

ACKNOWLEDGEMENTS

I want to thank a set of people who always supported me during my Ph.D. study. First of all, my supervisor Prof. Dr. Mine Elif Karsligil supported me with her invaluable guidance and prescience. During this study, I learned a lot of things about the science and the real life.

Besides that, I would like to thank my family; My father, my mother, my little sister, and my wife. I really owe you a lot during this study. Sometimes we missed some special days, sometimes we couldn't be together. For everything, besides my thanks, I also would like to present my apologies. But, I really can't imagine what would be if I didn't have you.

Last but not least, a Ph.D. study is not just a set of courses and exams, etc. It's a completely new perception sense of our world. To build this perception, I had discussions, listened to invaluable experts so on and so forth. I have to admit that, it's so hard to give some names to thank for the construction of this new sensation. For this reason, I would like to thank everyone who talked, criticized, and asked in other words had any interaction with me.

Erol ÇITAK

TABLE OF CONTENTS

LIST OF ABBREVIATIONS vii

LIST OF FIGURES viii

LIST OF TABLES x

ABSTRACT xii

ÖZET xiv

1 INTRODUCTION 1
- 1.1 Problem Definition . . . 4
- 1.2 Outline . . . 7

2 LITERATURE REVIEW 9
- 2.1 Datasets . . . 10
  - 2.1.1 Image Domain Datasets . . . 10
  - 2.1.2 Video Domain Datasets . . . 12
- 2.2 Related Works . . . 14
  - 2.2.1 FSL and ZSL for Image Domain . . . 14
  - 2.2.2 FSL and ZSL for Video Domain . . . 17

3 A NEW MULTI-MODAL VIDEO CLASSIFICATION DATASET 20
- 3.1 Why Do We Need a New Dataset? . . . 20
- 3.2 HVU-LD Dataset Construction . . . 25
  - 3.2.1 Class-name Preprocessing . . . 26
  - 3.2.2 Semantic Feature Extraction . . . 27
- 3.3 Dataset Summary . . . 28

4 A SEMI-SUPERVISED VARIATIONAL AUTOENCODER FOR FEW-SHOT AND ZERO-SHOT LEARNING SETTINGS 30
- 4.1 Motivation and Problem Definition . . . 30
- 4.2 Proposed Method . . . 31
  - 4.2.1 Frame Selection and Visual Feature Extraction . . . 32

4.2.2 Textual Information Pre-processing and Textual Feature Extraction . . . 32
4.2.3 Variational Inference & Autoencoder . . . 33
4.2.4 Cross and Distribution Alignment . . . 34
4.2.5 NMC: Non-Parametric Metric-Learning Classifier Module . . . 36
4.2.6 SS-CADA-VAE Loss . . . 38
4.2.7 Implementation Details . . . 38

5 EXPERIMENTS **41**
5.1 Few-Shot Classification Results . . . 43
5.1.1 Impact of Visual Feature Methodologies and Distribution Distance-Based Classification Module for FSL . . . 45
5.1.2 Impact of the Number of Classes in the Episode and Distribution Enhancement Module for FSL . . . 46
5.1.3 Impact of the DDC Module's Coefficient for FSL . . . 47
5.2 Zero-Shot Classification Results . . . 49
5.3 Few-Shot Classification on Existing Methods and Datasets . . . 50
5.4 FSL and ZSL Classification Final Remarks . . . 52

6 CONCLUSION **57**
6.1 Future Works . . . 58

7 APPENDIX **59**

REFERENCES **60**

PUBLICATIONS FROM THE THESIS **67**

LIST OF ABBREVIATIONS

FSL	Few-shot Learning
ZSL	Zero-shot Learning
HVU	Holistic Video Understanding
HVU-LD	Holistic Video Understanding Low-Data
SS-CADA-VAE	Semi-supervised Cross- and Distribution Alignment Variational Autoencoder

LIST OF FIGURES

Figure 1.1 A visualization of a metric-learning-based model's objective (The frames are taken from the Holistic Video Understanding dataset). 5

Figure 1.2 An episode of a meta-learning testing stage with support and query sets components besides the visual and textual information. In this illustration, we are considering the 3-way 3-shot learning setting, which indicates the existing of three different classes while each class has three different samples both in support and query sets. While the illustration classes are taken from our meta-test set, it's crucial to know that, meta-train, meta-validation, and meta-test sets share the same support and query sets structure (The frames are taken from the Holistic Video Understanding dataset). 8

Figure 3.1 A visualization of meta-sets in UCF101 dataset. 22

Figure 3.2 A visualization of meta-sets in HMDB-51 dataset. 23

Figure 3.3 A visualization of meta-sets in Kinetics-400 dataset. 23

Figure 3.4 A visualization of meta-sets in Something-2-Someting v2 dataset. 24

Figure 4.1 SS-CADA-VAE Model with Semi-Supervision module (SSM). SSM includes two regularization terms to enhance visual distribution parameters in terms of decreasing intra-class distance and increasing inter-class distance. In addition to regularization, SSM includes a distribution similarity-based classification operation which is used in both the few-shot setting and the zero-shot setting. 31

Figure 4.2 The coefficients of the loss objectives in SS-CADA-VAE are presented with their initial and finalized values, and increment strategy (Semi-Supervision objective aka $\mathcal{L}_{SS}$, has been experimented with different values, which are presented in the experiments section). 40

Figure 5.1 The system architecture of SS-CADA-VAE in the FSL and ZSL test classification stages. Both stages share the common distribution construction step, but then they differ in the context of the modalities used during classification. 42
Figure 5.2 Visual analysis of the impact of the DDC module on the HVU-LD meta-test 49
Figure 5.3 The confusion matrix of FSL for 45k episodes. (Best viewed in colored and zoomed-in.) 53
Figure 5.4 A visual example between 'bending back' and 'headstand' classes. (Best viewed in colored and zoomed-in.) 54
Figure 5.5 The confusion matrix of ZSL for 45k episodes. (Best viewed in colored and zoomed-in.) 55
Figure 5.6 A visual example of misclassification between 'feeding goats' and 'shearing sheep' classes. (Best viewed in colored and zoomed-in.) 56
Figure 5.7 Numerical representations of the unique and intersected textual descriptions of feeding goats and shearing sheep classes. 56

LIST OF TABLES

Table 1.1 Stats of some deep neural network models in terms of their number of parameters and sizes in MB (values are calculated by taking the original papers into account and by hand with 32-bit precision. Different precision strategies may impact the results). . 2

Table 2.1 Stats of low-data-oriented datasets for the image domain 10

Table 2.2 Stats of low-data-oriented datasets for the video domain 12

Table 2.3 The overall distribution of the Holistic Video Understanding dataset per category number of labels and the number of videos. . 14

Table 3.1 Cross-test distance of each meta-set based on the 5 most similar samples. The scores are between 0.0 and 1.0 and represent the average distance of the 5 most similar samples in cross meta-sets. (Higher is better) 25

Table 3.2 Distance analysis of the meta-training, meta-validation, and meta-testing classes in terms of Euclidean distance. 28

Table 4.1 An example of normalization over some textual inputs. The cumulative transformation of each sample input text is presented in each line 33

Table 5.1 Results of *5-way* few-shot video classification on the meta-test set. (Results are presented in percentage). 46

Table 5.2 Results of *1-shot* few-shot video classification on the meta-test set for different *n-way* settings. (Results are presented in percentage). 47

Table 5.3 Results of *5-way* few-shot video classification on the meta-test set for different γ coefficient values on CNN-based architecture with different k-shot values. (Results are presented in percentage). . . 48

Table 5.4 Results of zero-shot video classification on the meta-test set for different visual and textual architectures (results are presented in percentage). 50

Table 5.5 Results of *5-way* few-shot video classification on the HMDB-51 meta-test set on ViT-based architecture with different k-shot values. (Results are presented in percentage.) "-" indicates that there is no value presented in the original paper for that experiment. 52

Table 7.1 Classes of the meta-sets of the HVU-LD low-data-oriented dataset 59

ABSTRACT

Video Classification And Retrieval With Low - Data Regime Learning

Erol ÇITAK

Department of Computer Engineering
Doctor of Philosophy Thesis

Supervisor: Prof. Dr. Mine Elif KARSLIGİL

Video classification is a challenging task in computer vision that requires analyzing the content of a video to assign it to one or more predefined c ategories. However, due to the vast amount of visual data contained in videos, the classification process is often computationally expensive and requires a significant a mount of annotated data. Because of these reasons, the low-data-based video classification area, which consists of few-shot and zero-shot tasks, is proposed as a potential solution to overcome traditional video classification-oriented c hallenges. Although low-data-based studies are considered potential solutions, existing datasets, which are either not diverse or have no auxiliary modality that is a mandatory requirement for the zero-shot task, don't fulfill t he r equirements f or f ew-shot a nd zero-shot tasks completely. To address this gap, in this paper, we propose a large-scale, general-purpose dataset for the problem of multi-modal low data-based video classification. T he d ataset c ontains p airs o f v ideos a nd a ttributes t hat capture multiple facets of the video content. Thus, the new proposed dataset will both enable the study of low-data-based video classification t asks a nd provide consistency in terms of comparing the evaluations of future studies in this field. Furthermore, to evaluate and provide a baseline for future works on our new proposed dataset, we present a variational autoencoder-based model that leverages the inherent correlation among different modalities to learn more informative representations. As well as we introduce a regularization technique to improve the baseline model's generalization performance in low-data scenarios. Our experimental results reveal that our proposed baseline model, with the aid of this

regularization technique, achieves over 12% improvement in classification accuracy compared to the pure baseline model with only a single labeled sample. Our dataset and proposed baseline model can be used in future studies as a first step in this area. As a result, a wide range of video-based applications, including surveillance, video summarization, and content-based video retrieval, can benefit from this vision.

Keywords: Multi-modal dataset and learning, few-shot learning, zero-shot learning, video classification, deep variational autoencoder

ÖZET

Düşük Veri Rejimi Öğrenimi ile Video Sınıflandırma ve Geri Alma

Erol ÇITAK

Bilgisayar Mühendisliği Anabilim Dalı
Doktora Tezi

Danışman: Prof. Dr. Mine Elif KARSLIGİL

Video sınıflandırma, bir videonun içeriğini analiz ederek onu bir veya daha fazla önceden tanımlanmış kategoriye atamayı gerektiren, bilgisayarlı görü alanında zorlu bir görevdir. Ancak, videolarda bulunan büyük miktardaki görsel veri nedeniyle, sınıflandırma süreci genellikle hesaplama açısından pahalıdır ve önemli miktarda etiketli veri gerektirir. Bu nedenlerden dolayı, az veri ile video sınıflandırma alanı, az örnekle öğrenme (few-shot learning) ve sıfır örnekle öğrenme (zero-shot learning) görevlerinden oluşan, geleneksel video sınıflandırma ile ilgili zorlukları aşmak için potansiyel bir çözüm olarak önerilmektedir. Az veri ile yapılan çalışmalar potansiyel çözümler olarak görülse de, mevcut veri setleri, ya çeşitli değil ya da sıfır örnekle öğrenme için zorunlu olan yardımcı modalliklere sahip değildir. Bu nedenle az örnekle ve sıfır örnekle öğrenme görevlerinin gereksinimlerini tam olarak karşılamamaktadır. Bu boşluğu gidermek için, bu çalışmada, çok modaliteli az veri ile video sınıflandırma problemi için büyük ölçekli, genel amaçlı bir veri kümesi öneriyoruz. veri kümesi, video içeriğinin birden çok yönünü yakalayan video ve öznitelik çiftlerini içermektedir. Böylece, yeni önerilen veri kümesi, az veri ile video sınıflandırma görevlerinin incelenmesini sağlarken, bu alandaki gelecekteki çalışmaların değerlendirmelerinin karşılaştırılmasında tutarlılık sağlayacaktır. Ayrıca, yeni önerilen veri kümesimizde gelecekteki çalışmaları değerlendirmek ve bir temel oluşturmak için, farklı modallikler arasındaki içsel korelasyonu kullanarak daha bilgilendirici temsiller öğrenen varyasyonel otokodlayıcı tabanlı bir model sunuyoruz. Bunun yanı sıra, temel modelimizin az veri senaryolarında genelleme performansını artırmak için

bir düzenleme tekniği öneriyoruz. Deneysel sonuçlarımız, bu düzenleme tekniğinin yardımıyla, önerilen temel modelimizin, yalnızca tek bir etiketli örnekle temel modele kıyasla sınıflandırma doğruluğunda %12'den fazla bir iyileşme sağladığını ortaya koymaktadır. veri kümesimiz ve önerilen temel modelimiz, bu alanda ilk adım olarak gelecekteki çalışmalarda kullanılabilir. Sonuç olarak, gözetim, video özetleme ve içerik tabanlı video arama gibi geniş bir yelpazedeki video tabanlı uygulamalar bu vizyondan faydalanabilir.

Anahtar Kelimeler: Çok modlu veri kümesi ve öğrenme, az örnekle öğrenme, örneklemsiz öğrenme, video sınıflandırma, derin varyasyonel otomatik kodlayıcı

1
INTRODUCTION

Deep learning-based statistical learning algorithms have shown superior results in various research topics and tasks. One of the first scientific investigations was published in 1943 by McCulloch and Pitt [1] about how neurons work and can be mathematically formulated. Later, Yann LeCun's research about the convolutional neural network [2] achieved a significant breakthrough for those years. Although all these works exist, because of the lack of technology and data, deep learning-based methods leaped forward in 2012 right after the ImageNet [3] challenge for image classification and object detection.

Deep learning-based algorithms have shed light on various research topics, especially since 2012. In this context, the developments in the field of the computer vision era are also getting more quickly with the help of pioneer study ([4]). In addition to the image recognition problem, object segmentation ([5]), 2D-3D scene reconstruction ([6]), image recognition with localization ([7]), image-video retrieval ([8]), disease prediction ([9]), biomedical verification ([10]) and many other topics have beaten their state-of-the-art performances with the help of deep learning and especially convolutional neural networks and more advanced optimization techniques ([11], [12]).

However, researchers in the era of natural language processing and understanding era have achieved enormous success with the help of deep learning. First of all, word and sub-word-based token embeddings are proposed to represent tokens in numeric vector form ([13], [14]). Apart from these researches, many different problems including named entity recognition ([15]), sentiment classification ([16]), and many other topics have beaten their -generally- TF-IDF and one-hot encoding-based state-of-art performances with the help of long-short term memory networks or Transformers ([17], [18]).

In addition to both computer vision and natural language processing eras, tabular data [19], audio processing [20], and multi-modal processing ([21], [22]) eras are

also other remarkable research topics.

Deep learning-based algorithms are generally trained with the first derivative-based iterative stochastic gradient descent algorithms. These algorithms require differentiable elements in each component of deep learning-based algorithms and follow different data sampling procedures over many data points for training. Due to its iterative nature -especially with the mini-batch sampling technique- and the total number of parameters in its network layers, deep learning-based algorithms require a large dataset for training. Table 1.1 shows some of the deep neural network algorithms and their statistics in terms of the number of layers, the number of parameters, and the sizes in MB.

Model Name	Size in MB	Parameters	Depth
Xception	88	22.9M	126
VGG19	549	143.6M	26
ResNet50	98	25.6M	50
InceptionV3	92	23.8M	159
MobileNet	16	4.2M	88
MobileNetV2	14	3.5M	88
DenseNet121	33	8M	121
EfficientNetB1-7	31 - 256	7.9M - 43.3M	-
NASNetLarge	343	88.9M	533
UNet	118	31M	10
BiSeNet	22	5.8M	-
YoloV9	27 - 220	7.2M - 58M	-
Capsule Networks	35	8.2M	-
TimeSformer	484	121M	-

Table 1.1 Stats of some deep neural network models in terms of their number of parameters and sizes in MB (values are calculated by taking the original papers into account and by hand with 32-bit precision. Different precision strategies may impact the results).

Deep learning-based algorithms for both image-processing and video-processing domains have been applied to many different problems, and only a small portion of them are presented in Table 1.1. In this context, as the proposed solutions are getting more complex, the possibility of experiencing of bias-variance trade-off increases [23]. At this juncture, the variety of a training dataset comes up as one of the exact solutions to alleviate this problem before different learning strategies.

In conclusion, deep learning-based algorithms are high data consumers to figure out the patterns that reside in the data. And also they require high regularization discipline to overcome bias-variance trade-offs successfully. Otherwise, training error or test set error will be too high which is called zero learning or memorizer

learning [23]. After realizing all these separately, if we take a step back and look at real-world learning conditions, the situation is quite different. Real-world scenarios demand high achievements with less labeled data. Also according to the YouTube stats, in every minute 500 hours of video were uploaded to YouTube in 2019. And the vast majority of the uploaded videos have one or zero tags [24]. Under this high volume of unlabelled data explosion, unsupervised and semi-supervised deep learning algorithms, which either require only a small amount of data or no labeled data, are gaining more attention.

Besides that, meta-learning, or learning-to-learn [25], is a learning policy that has a group of tasks, a training procedure for those tasks, and a performance monitor to improve the learning model's generalization capacity in each task systematically and iteratively.

In light of the recent growing unlabelled enormous size of video data, with the cost of data labeling for video classification tasks, due to the vast amount of visual data contained in videos, our study will be investigating the few-shot and zero-shot meta-learning policies for video classification problem. In this context, our study outcomes can be listed as follows;

- First, we propose a dataset due to current video classification datasets are being highly biased towards specific actions, which have low diversity, and lack auxiliary information beyond class labels to enable proper zero-shot learning and multi-modal learning. Our dataset is based on the Holistic Video Understanding dataset [26], which we have adopted for meta-learning-oriented low-data-based video classification problems. To accomplish this, we propose a dataset called Holistic Video Understanding over Low Data (HVU-LD), which utilizes the meta-learning paradigm to enable episodic learning. The dataset is divided into three sets - meta-training, meta-validation, and meta-testing - each with unique classes that do not appear in the other sets. The Training, validation, and testing stages are applied to the meta-training, meta-validation, and meta-testing sets, respectively. Each set is composed of multiple episodes, each of which includes *n-way k-shot* support-query pairs. In the meta-learning paradigm, the support set is called the reference set, and the query set is called the evaluation set. The n refers to the number of classes in the episode, and k refers to the number of samples per class in the support set.
- Second, in addition to proposing a multi-modal low-data-based video classification dataset, we have analyzed our proposed dataset and existing

video classification datasets, which are proposed for video classification and then adapted to low-data-based learning even if have not enough contextual information, in terms of semantic distance. Using this semantic distance, we aimed to create a low-data-based dataset that closely reflects real-world conditions. Our goal was to ensure that not only do the meta-sets contain different classes, but also that the classes in the meta-sets are semantically distant from each other.

- Third, we prefer to choose a supervised variational autoencoder model that applies to both few-shot learning and zero-shot learning scenarios as a baseline method to evaluate this new dataset. This baseline model uses the CADA-VAE [27], which was introduced for generalized few-shot and zero-shot learning in the image classification domain. Since the baseline model was developed for the still-image domain, minor changes have been made to make it available for the video domain. This model comprises two variational autoencoder networks for each modality, incorporating cross-alignment and distribution alignment objectives.

- Lastly, we extend this baseline, CADA-VAE, model by a distribution-based regularization technique to enhance the generative capability of the variational autoencoders while increasing the inter-class distance and decreasing the intra-class distance. The proposed modification to the baseline model architecture enables supervised training while maintaining its efficacy for few-shot and zero-shot learning. This modification incorporates meta-learning techniques that allow the model to learn how to quickly adapt to new tasks with limited examples. By training the model through episodic training, it learns to generalize better from limited examples, resulting in improved performance for few-shot and zero-shot learning scenarios.

1.1 Problem Definition

The importance of data labeling is considered a crucial step in supervised learning algorithms' performances. Within this context, several different solutions have been proposed like a hand-made data labeling platform like Amazon Mechanical Turk (MTurk)[28], or a semi-supervised (semi-automated) data labeling algorithm like active learning [29], etc.

Besides the data-centric approaches, low-data-oriented learning policies have been proposed, which are capable of dealing with bias-variance trade-offs while the data consumption demand is considerably low. During our study, we will be investing

in the meta-learning, or learning-to-learn, approach for video classification and retrieval problems.

Meta-learning consists of three main steps. The first one is distributed tasks, the next one is a performance measurer for learning optimization, and the last one is a learning algorithm to wrap the first two components together to achieve the main goal. In this study, our main goal is video classification & retrieval with the metric-based meta-learning policy.

The metric-based meta-learning policy aims to adjust the distance between sample data points according to their class information. In this sense, the data points from the same class distribution are expected to be as close as possible in the embedding space, and for the different class distribution-based samples, the total opposite of this behavior is desired. In Figure 1.1, the real-world data samples belong to three different classes; riding a bike, rock climbing, and eating cake. Metric learning-based models' objective is to map the input data samples, the video frames in our study, into an embedding space and find a decision boundary between those samples while considering their class information.

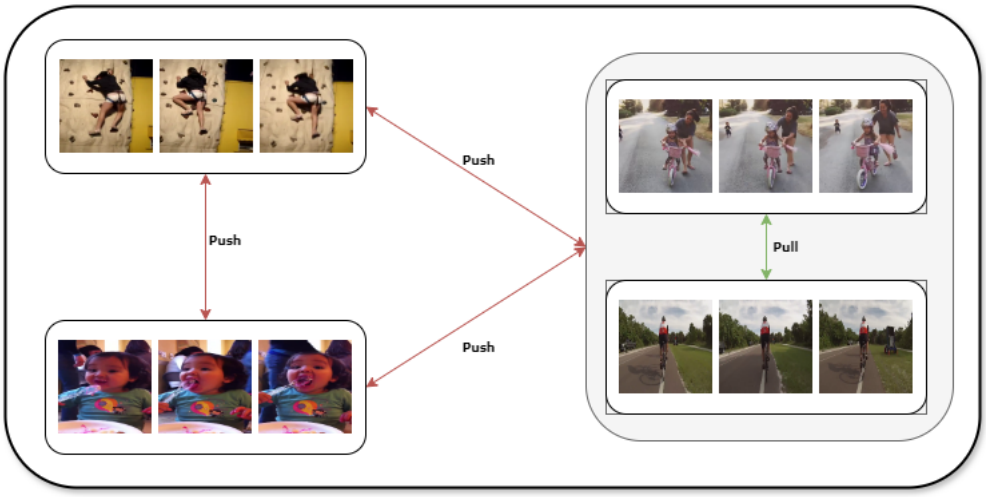


Figure 1.1 A visualization of a metric-learning-based model's objective (The frames are taken from the Holistic Video Understanding dataset).

From this point forth, in the low-data-based learning scenario, learning models correspond to the situation where only a scarce number of samples are available during the training stage, which is a very common situation in the real world. In addition to this scarcity, there is another scenario without the training sample, which forces learning models to acquire the skills of comparison and the cause-effect relationship between different modalities. These two different low-data-based learning scenarios define key subdivisions; few-shot learning and zero-shot learning.

More precisely, In this illustration, we are considering the 1-shot, 5-class classification task where for each dataset, learning models are expected to learn new concepts from a few examples, which may lead the state-of-the-art deep learning models to face generalization problems. In addition, in the zero-shot learning scenario, which is a more challenging task, no examples are presented for new concepts to the learning models during the training stage and the concepts without samples are expected to be predicted using additional information from other modalities.

The junction point of few-shot and zero-shot learning approaches with meta-learning can be expressed as a *N-way K-shot* learning setting. In this way, the dataset is split into three pieces; meta-training, meta-validation, and meta-testing. Unlike the conventional supervised algorithm policy, these sets have unique classes, in other words don't contain joint data classes. After obtaining the meta sets, then each set is used in *N-way K-shot* setting. Here *N* refers to the number of classes in each task and *K* refers to the number of samples for each class. This organization is also called as a *episodic learning*, and, follows one of the main components of meta-learning, distributed tasks-based learning. This episodic learning setting is presented in Figure 1.2. In this learning setting, the classes and the samples are determined with the aforementioned policy. Its iterative distributed task-based learning mechanism is done *Support - Query Sets*, which corresponds to the training and testing samples respectively. Regardless of the meta-set, each meta-set consists of a support set for training and a query set for testing purposes.

As a conclusion, in the zero-shot video classification task, the model is expected to classify the visual modality samples -videos- in the query set using the auxiliary information -video level semantic labels- in the support set. The task definition of ZSL is expressed as follows. Let's define the meta-training set as follows;

$$D_s = \{(x, y, a(y)) \mid x \in X^s, y \in Y^s, a(y) \in A^s\} \tag{1.1}$$

, where consisting of visual features *x*, class labels *y*, and auxiliary information *a(y)*. In this equation *s* denotes the seen classes. At the same time, the meta-testing dataset can be defined as follows;

$$D_u = \{(x, y, a(y)) \mid x \in X^u, y \in Y^u, a(y) \in A^u\} \tag{1.2}$$

, where *u* denotes the unseen classes and the intersection of Y_s an Y_u is a null set.

In the inference stage, the learning model is aimed to obtain good relations between $A^u \rightarrow Y^u$ for the query set visual modality samples, X^u.

Similar to the zero-shot setting, few-shot learning includes the D_s and D_u meta sets, but these sets do not have any auxiliary information, and in the meta-testing phase, the learning model is expected to learn the relationship between $X^u \rightarrow Y^u$ using visual samples in the support and query sets.

After analyzing Few-shot learning (FSL) and Zero-shot learning (ZSL) problems, in terms of their data structure, etc., there is still one blurry point is needed to be clarified. In the literature, there are two different types of FSL and ZSL problems. The first type is called *conventional FSL and ZSL* and the other one is *generalized FSL and ZSL* [27]. Even if this thesis focuses on the conventional side, their difference is required to be known. Conventional FSL/ZSL problem refers that the images to be recognized at test time belong only to novel/unseen classes. The generalized FSL/ZSL problem refers to the images to be recognized at test time belong to base and novel, which means seen and unseen classes. While the generalized version of FSL and ZSL problems are more realistic and suitable for real-world conditions, we first focused on closing the gap in conventional FSL and ZSL problems in the video domain, then our future work will extend the current thesis to generalized FSL and ZSL approach.

1.2 Outline

The organizational structure of this thesis can be presented as follows; literature review for FSL and ZSL problems will be presented in Section 2. More precisely, both image and video-related FSL and ZSL algorithms will be analyzed in subsections to understand the historical background and also the connection between the two domains in the FSL and the ZSL concept. The dataset construction and model proposal with its deep experiments will be presented in Section 3. In the last section, Section 4, all the thesis will be concluded and future works in this domain will be shared.

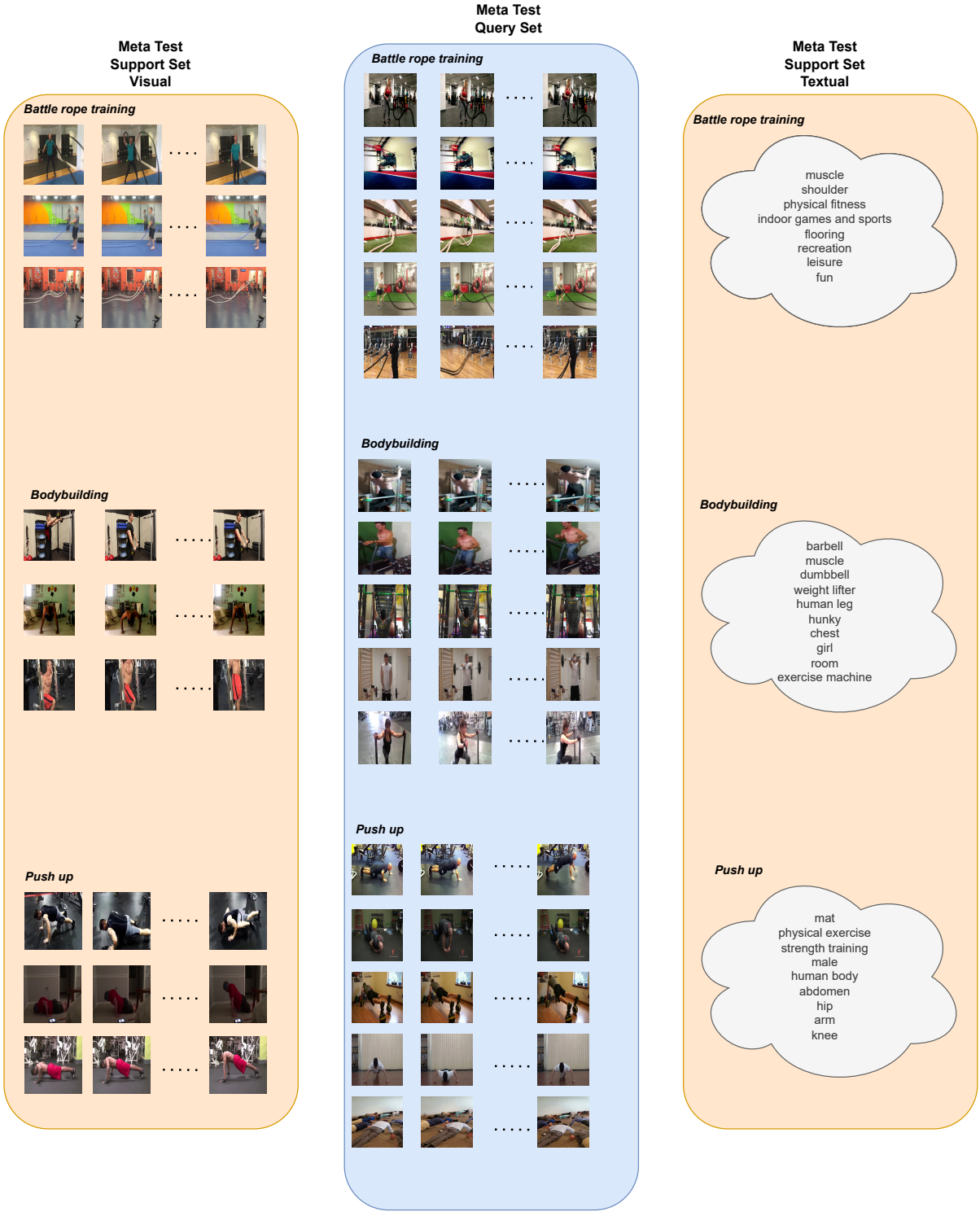


Figure 1.2 An episode of a meta-learning testing stage with support and query sets components besides the visual and textual information. In this illustration, we are considering the 3-way 3-shot learning setting, which indicates the existing of three different classes while each class has three different samples both in support and query sets. While the illustration classes are taken from our meta-test set, it's crucial to know that, meta-train, meta-validation, and meta-test sets share the same support and query sets structure (The frames are taken from the Holistic Video Understanding dataset).

2
LITERATURE REVIEW

In this section, we will analyze the analogy of the FSL and ZSL algorithms in two different groups. While the first group will be related to the still-image domain, the second group will analyze the video domain. Besides the state-of-the-art algorithms, the datasets for each group will also be analyzed separately.

But, in general, FSL and ZSL models, or low-data-based learning models to indicate two learning models at the same time with a framework term, consist of three primary different perspectives to alleviate the bias-variance trade-off, which is caused by the low number of data.

The most intuitive strategy is data augmentation which aims to increase the number of samples while considering the prior knowledge by using hand-crafted techniques like; flipping, color changes, rotation changes [30], or novel learning model-based data hallucination models [31, 32].

Another strategy to alleviate the low-data-oriented overfitting issue is using complex hypothesis constraint strategies which try to constraint the search space while considering the prior knowledge. Embedding learning-based approaches [33–35], like task-based learning techniques or parameter-sharing oriented multi-task learning-based studies [36] can be given as examples.

The final strategy takes prior knowledge and makes a composition to find out the best optimizer step for the optimal parameter group. Either fine-tuning the existing parameters by robust regularization approaches like [37], which cooperate with early stopping, $l*$-norms, etc., or meta-learned parameter approaches like [38], which use a couple of first-derivative based optimizer steps can be given as examples for this strategy group.

2.1 Datasets

In this section, datasets for low-data-oriented approaches will be analyzed in two groups; image domain and video domain-based ones. Dataset analysis carries too much importance to show the efficiency and performance of the models. Also, it helps to understand the datasets' abilities to cover real-world conditions. Because of these factors, the datasets will be analyzed while taking the following topics into consideration;

- number of classes
- number of samples in each class
- train, test, and validation split policy
- semantic distance between classes in each split
- distribution of classes in terms of variability
- additional context that the dataset includes

Those topics will provide information about a dataset of its suitability in real-world situations, validity for FSL and ZSL problems, degree of difficulty for low-data-oriented models, etc. In this context, the following subsections will explore the datasets for their domains while considering the above bullet points.

2.1.1 Image Domain Datasets

In the image domain, there are several options for FSL and ZSL problems. The summarized information can be obtained from Table 2.1.

Dataset Name	Number of Classes	Number of Samples in each Class	Additional Context
Omniglot [39]	1623	20	No
miniImagenet [40]	100	600	No
CUB [41]	200	31	Yes
SUN [42]	>700	∼200	Yes
AwA1 [43]	50	∼610	Yes
AwA2 [44]	50	∼710	Yes

Table 2.1 Stats of low-data-oriented datasets for the image domain

Omniglot dataset [39] includes 1623 classes of handwritten characters from 50 different alphabets. In the base study, 30 alphabets were split for learning stages

(training, validation), and 20 alphabets were used for the evaluation stage. No study has applied semantic distance analysis between alphabets in the learning and evaluation stages before. Besides that, it's clearly seen that this dataset is only valid for few-shot learning-based problems because of the lack of additional context in the alphabets.

Imagenet dataset [45] is a famous dataset for image classification and object detection domains due to its support for the development of convolutional neural networks like Alexnet [4] in 2012. The researchers in [40] proposed to use this dataset in few-shot-learning-based problems. From the main Imagenet dataset, 100 classes were selected randomly. Even if the original dataset has enough amount of randomness in other terms good variability, the output of this random selection was not analyzed by the authors. In total, the number of classes for training, validation, and testing is 64,12,24 respectively. Also, there is no additional context for the images in the original dataset. Because of this property, this dataset is also only valid for shot-learning-based problems.

Apart from the FSL-oriented datasets, CUB [41], SUN [42], Awa1 [43], and Awa2 [44] are the datasets for both FSL and ZSL problems due to their additional context option. Since the ZSL problems require an additional modality besides the visual data, those datasets have been adopted and used for FSL and ZSL problems as benchmark datasets. Within those datasets, while CUB, AwA1, and Awa2 are specialized for the animals, SUN is a dataset that focuses on scene and scene attributes.

CUB focuses on birds with fine-detailed subcategories and this feature makes this dataset challenging for not just FSL or ZSL problems but also traditional supervised classification problems. Because the variances are getting smaller in subcategories.

On the other hand, the AwA1 and AwA2 datasets focus on animals in a more broader view, with their specific attributes like color, habitat, and eating behavior. Even if AwA2 especially nearly has $\sim$85 different attributes for an animal, the number of classes in the dataset, which equals 50, seems like a critical limitation to make an evaluation in a broad sense.

Except for those datasets, there are also some other datasets, even if they have been used less frequently. Cifar-FS [46], Stanford Dogs [47], TieredImageNet [45], and Stanford Cars [48] are examples of less frequent datasets in the FSL problem.

2.1.2 Video Domain Datasets

In this subsection, the video domain datasets that have been used for FSL and ZSL problems will be analyzed in deep while considering the points, which are presented in Section 2.1. Since this thesis proposes a dataset for FSL and ZSL-based video classification problems, the analysis will pay more attention to existing datasets by these aspects; the coverage of real-world cases and the distribution of classes for each set, i.e., training, validation, and test sets' classes.

The analysis will start with presenting the existing datasets and their properties compactly. In this context the summarized information about video domain datasets can be obtained from Table 2.2.

Dataset Name	Number of Classes	Number of Samples (in Total)	Additional Context
UCF-101 [49]	101	13.320	No
HMDB-51 [50]	51	6.766	No
Kinetics-400 [51]	400	306.245	No
ActivityNet [52]	200	10.024	Yes
Something-2-Something [53]	174	220.847	Yes
HowTo100M [54]	23k	136.6M	Yes
Holistic Video Understanding [26]	3K	577K	Yes

Table 2.2 Stats of low-data-oriented datasets for the video domain

According to Table 2.2, some datasets do not include any additional context, in other terms another modality than visual modality. This state limits their usage in the ZSL problem by itself.

In this context, UCF-101 dataset includes 101 different classes of actions. The dataset includes 5 different action categories: human-object interaction, body-motion only, human-human interaction, playing musical instruments, and sports. Although the dataset, which offers approximately 130 videos in each class, is not suitable for ZSL problems because it does not have additional context, it is perfectly suitable for FSL problems as it has rich content in terms of diversity.

HMDB-51, a large human motion database, is a video action recognition dataset that focuses on human motions. Even though it may seem to have a very limited structure at first glance, because of its name, this dataset can be described as a diverse dataset in terms of the distribution of classes it contains, such as laughing, kicking a ball, mountain climbing, etc. It has nearly 130 videos per class and no additional context information in itself.

The Kinetics-400 dataset comprises around 400 classes of human actions, ranging from common activities to specialized motions. With over 300,000 video clips, it offers a diverse representation of human actions in various contexts. The actions are human-focused but the dataset's value is more about the richness and complexity of human activity in various settings.

The ActivityNet dataset is similarly designed to capture a broad spectrum of human activities, ranging from everyday actions to more specialized tasks. With its extensive collection of temporally annotated videos spanning various durations, it provides a comprehensive representation of real-world activities. While it includes specific categories like sports and household chores, its diversity mirrors the complexity of human behavior across different environments and scenarios. Since each segment of the videos is annotated by humans from Amazon Mechanical Turk, it is suitable for both FSL and ZSL-based problems. Regardless of its suitability, the annotations in ActivityNet primarily focus on temporal boundaries and class labels, lacking detailed contextual information about the scenes, objects, or interactions depicted in the videos.

Something-2-Something dataset features a wide array of fine-grained human actions, totaling over 220,000 videos across 174 action classes. With its rich temporal annotations, it offers precise delineation of actions within the videos, aiding in action localization tasks. Even though the class distribution itself does not guarantee good real-world coverage, the dataset has a major advantage in that it contains videos of people performing basic daily actions with everyday objects.

HowTo100M dataset includes videos from a video-sharing platform, YouTube, about instructional videos. As its name indicates, the videos are related to answering the *How to* questions. It includes 23611 different tasks and more than 136M video clips. The video topics are determined from wikiHow [55], a Wiki-based instructions website, and applied extensive filtering to eliminate irrelevant videos. But both the low human-controlled labeling, only automatic speech recognition based on additional context, and lack of scene boundaries make this dataset harder and more noisy for low-data-oriented problems.

The Holistic Video Understanding (HVU) dataset is a significant contribution to the field of video understanding, as it addresses the limitations of existing video classification datasets that mainly focus on human actions or sports recognition without providing detailed information on the entire video. With approximately 572k videos and 9 million annotations covering 3142 labels across 6 semantic categories, including scenes, objects, actions, events, attributes, and concepts, HVU

offers a large-scale and diverse dataset for video classification research.

To label the 6 semantic categories, a two-stage framework was employed that involved both machine-generated automatic labeling and human-based validation processes. The Google Vision API [56] and Sensifai Video Tagging API [57] were used for the machine-generated labeling stage, and a human validation stage was performed to eliminate any erroneous labels. This framework allowed for the combination of both processes, resulting in accurate labeling. After that two-stage labeling process, the statistics including the number of labels and videos for each category are presented in Table 2.3. Each row indicates the category name, the number of distinct labels per category, and the number of videos that have that category-oriented label.

Category	Number of Labels	Number of Videos
Scene	419	366,941
Object	2651	480,821
Action	877	481,418
Event	149	320,428
Attribute	160	369,668
Concept	122	375,664

Table 2.3 The overall distribution of the Holistic Video Understanding dataset per category number of labels and the number of videos.

2.2 Related Works

In this section, the related works for low-data-oriented problems will be based on their domains; image domain and video domain. In each subsection, the previously proposed studies will be analyzed from both the FSL perspective and the ZSL perspective. Besides that, as introduced before, several approaches have been proposed to handle FSL and ZSL problems. But in this thesis, our approach will focus on the metric-learning-based methods. However, since we will be proposing a dataset for the low-data-based video classification problem, any method that either follows the metric-based learning approach or other approaches can benefit from it. For this reason, even if our approach is metric-based learning, all significant studies will be presented in respective subsections.

2.2.1 FSL and ZSL for Image Domain

The first approach will involve increasing the number of data samples to mitigate the negative effects of limited data during the learning phase. In this way,

the methods either propose naive data augmentation operations like translation, flipping, rotation, and shearing or use more complex learnable approaches.

Even if naive data augmentation techniques are easy to use for data augmentation, the main goal of data augmentation is not satisfied well with this type of data augmentation. Since these hand-crafted operations are not specific to the given problems and suffer to increase the inter-class variability low data regime learning algorithms are required to have more sophisticated data augmentation techniques. In this context, [58] proposes a saliency-guided data hallucination method that finds the background and foreground parts of a given image. Then the combination of background and foreground parts, which are taken from different images, constructs the new augmented images. But to make more realistic augmented images, they designed a semantic similarity mechanism that measures the probability of harmony between classes of given images, which are first split into parts and combined to make an augmented image. In this way, incompatible pairs are discarded. Another data augmentation method, combining hand-crafted transformations with a feature embedding-based approach, is proposed in [59]. It constructs global class representations by averaging feature embeddings of base classes and employing a two-stage synthesis strategy for novel classes, involving random cropping, flipping, and data hallucination technique [60]. Synthetic data generation enhances intra-class variance, contributing to a more stable global class representation. Episodic-level class representations are compared with global ones, followed by registration loss computation and performance evaluation using mean squared error. On the other hand, some complex data augmentation approaches include generative networks. For example, [61] uses conditional Wasserstein Generative Adversarial networks, the study aims to regulate both new image features and a few-shot classifier at the same time.

Another important way to handle low-data-oriented problems is model-based learning. Model-based learning addresses the low data regime problem by learning to compare idea. This idea tries to constrain hypothesis space by prior knowledge. Thus even if the problem relates to the low data regime, the optimal solution can be reachable using this prior knowledge by constraining. Within the scope of constraining search space, multi-task learning, embedding learning, and generative learning-based works are proposed. The most popular of these works is embedding-based ones which are identified as metric-based learning. In this way, the samples are embedded into a lower-dimensional space then while the similar samples are expected to be close together and the dissimilar ones are separated. The first work that belongs to embedding-based learning is [62], which was published in 2016 by Google DeepMind. The study uses a large-scale meta-training set that has

been randomly sampled for each episode. And, each episode follows the *N-way-K-shot* which refers to the select N different classes and K different samples for each class. Each support set and query set image is fed to a 4-layered convolutional neural network that produces feature embeddings. Then using cosine similarity over the feature embeddings, the classification result is determined using softmax over the cosine similarities. In the end, categorical multi-class cross-entropy loss on the output probabilities of the softmax layer is backpropagated through the 4-layered convolutional neural networks to obtain better features. The matching network tries to compare images through their feature embeddings without any prior knowledge and this comparison helps to improve the capability and explainability power of the 4-layered convolutional neural network. Besides that, there are some other studies that have strong relations with matching networks. For example, ProtoNetworks [63] proposed two different novelties in the previous study, using K samples for prototype construction and using the Euclidean metric as a distance operator. Then after, [64] proposed a completely learnable distance operator instead of a pre-defined one like Euclidean, Cosine, etc.

In zero-shot learning, unlike few-shot learning, model evaluation is operated over the cross-modalities. More precisely, in zero-shot learning there are no samples are given in testing time and expected to infer similar class-oriented information between different modalities for each sample. This constraint implies that learning models have a shared feature space between different modalities. In this shared feature space, different modalities are enforced to have a joint representation for zero-shot evaluation. In this manner, [65] constructs the class prototypes using label information as an attribute. Or [66] proposes the representation transfer mechanism from a classifier-based training onto Prototype Networks [34]. Also, [67] proposed a multimodal prototypical network for image classification. In this multimodal case, an image is fed into an image encoder which consists of multi-convolutional layers, also the image's description in the form of natural language is fed into a text encoder. This text encoder encodes free-format text into feature embedding as in the image encoder. Then using a conventional generative adversarial network both class label and their scores of realness or fakeness are measured. When the training is done, the text encoder provides additional feature embedding to the image-based feature embeddings to construct more robust and stable class prototypes. On the other hand, the generative side, [68], generates visual and textual features with the help of autoencoders, which have two branches for each modality, for classification purposes, then minimizes the maximum mean discrepancy between visual and textual modality feature space distributions to obtain joint latent space feature. Alongside the latent distribution alignment, [27]

implements a cross-reconstruction loss which follows that each encoded feature for all modalities is decoded in all modalities decoders. In [69] implements the attribute-guided transformer mechanism for visual modality then in the decoding stage of this transformer language model-based textual feature is embedded into the visual features to obtain the latent space of the visual-semantic joint.

2.2.2 FSL and ZSL for Video Domain

The history of low-data-oriented video classification starts with [70]. In this study, authors proposed a dataset for FSL to close the gap in this research area using Kinetics dataset [51]. This dataset consists of meta-training, meta-validation, and meta-testing sets with 64, 12, and 24 classes respectively. And, each class is randomly selected from the Kinetics dataset, with 100 videos in each class. One of the biggest drawbacks of this random class-choosing approach is related to class distributions of meta-sets. In the conventional FSL and ZSL problems, the intersection of train and test sets is an empty set. Because of this reason, it's good to have as different as possible different distributions are desired in those meta-sets instead of randomly choosing. Also, [70] implements a two-stage architecture where the first stage is a visual descriptor with multi-specificity support, and the second stage is a key-value-driven memory network for few-shot learning. With this newly proposed dataset and Something-V2 [53] dataset, TAM [71] proposed a novel long-term temporal ordering information-based comparison module with the help of DTW algorithm [72] on top of embedded visual features of both support and query set videos. Through temporal alignment distance, TAM accomplishes the classification purpose of the few-shot learning scenario. [73] proposes a metric-learning approach with using the pre-trained Resnet-50 convolutional network. After getting the initial embedding with that network, the study uses a transformer-based embedding learning module with the help of the instance normalization module, which is employed to increase the discriminative power of the features. Thus far, many studies used highly popular convolutional neural networks, which were trained with the Imagenet dataset. But [74] proposes a solution for the few-shot action recognition problem with using the textual and visual encoders from CLIP [75] study. While the study keeps frozen the textual encoder, it slightly fine-tunes the visual encoder of CLIP. Since this model, like before ones in this video domain section uses datasets that have no additional context, it uses the tag of each video to create a synthetic textual information for the video. This textual information follows a specific textual structure, which is proposed by CLIP, like *"A photo of* $< video_label >$*"* structure.

Also, some other studies generate the syntactic textual information for the datasets

that suffer from additional context. In this context, [76] addresses to lack of a dataset with additional context in the few-shot video action recognition problem and proposes a solution with a foundation model BLIP [77]. The study generates the textual information for each video based on their frames using the pre-trained BLIP model's visual encoder and caption decoder. Then using a Transformer-Bert-based textual encoder and fine-tuned BLIP visual encoder outputs are projected onto a shared space to perform metric-based classification.

On the other hand, even if some approaches follow to use of 2D-based visual encoders, which are based on the frame level feature of the entire video with a spatial or temporal aggregation, [78] employs a 3D feature extractor with efficient-fusion model to perform metric-based classification. Besides the spatiotemporal analysis novelty, this paper is one of the first papers that employs ChatGPT3.5 for textual information generation.

Also, even if our study focuses on the conventional FSL and ZSL problems, there are some studies that handle this problem. For example, Generalized Many-Way Few-Shot Video Classification (GFSV) [79] proposed a 3D-CNN baseline in the feature extraction stage to capture spatiotemporal information. In this study, additional video samples were collected to overcome low data-oriented problems in FSL. In concrete, the tags for video-based search are applied to the *YFCC100M* [80] dataset. In the search process, Fasttext embedding method [14] [14] is used to compare which video samples are close to each other. Another crucial point of this study is the classification section which includes conventional FSL and generalized FSL predictions.

In zero-shot learning, the video domain has taken limited attention and requires a dataset that has a video-video description pair-based dataset. Until this time, the existing datasets belong to the specific and mostly human-based actions; UCF101 [49], HMDB51 [50]. They also do not contain a detailed description of the video such as scene, object, action, event, or attributes that are crucial for the real-world use cases. Nevertheless, [81] proposed an end-2-end architecture that uses Word2Vec word embeddings [13] to extract continuous embedding from the action class label, and unlike the previous works in zero-shot video classification, it implements trainable 3D-CNN for the visual feature extraction. And, the mean square error is used to project these two modality features into the shared feature space. Besides that, [82] proposes a cross-modal transformer network that can extract visual-semantic joint space representation simultaneously. In that study, the input of the cross-modal transformer network is organized like masked language models. Visual frames and textual information are fed into the model at the same

time with a classification token. The model is trained with two different objectives, The first one is the classification, which is based on the classification token and the second objective is the language prediction objective that first masks one of the input textual tokens in the beginning and tries to predict that token as an output of the transformer network.

3
A NEW MULTI-MODAL VIDEO CLASSIFICATION DATASET

The dataset choice is a crucial step in any statistical learning method. The learning curve, the generalization capacity, and many other parameters rely on this choice. In this context, this section will analyze both the previously existing datasets and our new proposed dataset for the low-data-oriented video classification problem. Even if the basic descriptive statistics of the previously existing datasets were examined in Section 2.1, their analysis based on the FSL and ZSL aspects was missed.

In these circumstances, the following subsections will first investigate the requirements of a new FSL and ZSL dataset. In this investigation, a detailed comparison between the previously proposed dataset and our proposed dataset will be done. Then, since the new proposed dataset will include textual context besides the visual context, the videos, the pre-processing methods, which are applied to the raw textual contents, will be explained in depth.

3.1 Why Do We Need a New Dataset?

The main objective of few-shot learning and zero-shot learning video classification paradigms is to facilitate the training of a learning model using the meta-training set while iteratively refining learning model parameters and hyper-parameters, such as learning rates, weight decay coefficients, and batch sizes, through validation on the meta-validation set. Upon completion of the meta-training and meta-validation phases, the model's performance in generalizing to unseen classes is assessed on the meta-testing set, which comprises novel classes. At this juncture, according to the meta-learning philosophy, the dataset must be split into three groups for meta-training, meta-validation, and meta-testing stages, while taking into account that each group's intersection must be disjoint.

From this main objective, in the zero-shot video classification task, the model is

expected to classify the visual modality samples -videos- in the query set using the auxiliary context -video level semantic labels- in the support set.

Similar to the zero-shot setting, few-shot learning includes the D_s 1.1 and D_u 1.2 meta sets, but these sets do not have any auxiliary context, and in the meta-testing phase, the learning model is expected to learn the relationship between $X^u \rightarrow Y^u$ using visual samples in the support and query sets.

At this point, zero-shot and few-shot learning settings need a dataset that strictly fits the following points to mimic real-world conditions;

- Additional context included videos to work on both zero-shot and few-shot learning settings
- Variable video classes to ensure real-world conditions - not being stuck in some specific field classes -
- High intra-class distances between meta-sets to test the generalization capacity of the learning model
- Low inter-class distances within meta-sets put pressure on the learning model to find out the fine-grained details

In this context, we first tested the previously proposed datasets in the light of these rules to validate their suitability and quality for zero-shot and few-shot learning settings. In this evaluation pipeline, besides whether a dataset includes an additional context or not, we also investigated the class distributions for each meta-set. This investigation states the dataset's compatibleness for real-world conditions. Also, since the semantically similar classes in different meta-sets may be interpreted as a data leakage. Even if the classes are not identical for these kinds of semantically similar classes, the classes may either include similar objects or similar movements, etc. like in *rope climbing* and *rock climbing*, which is a condition in UCF-101 based FSL and ZSL dataset split. At this point, detailed information about the pre-processing steps when creating semantic distances and the semantic vector acquisition method are presented in the section 3.2.

Firstly, the UCF101 dataset can be divided into 5 different main categories in terms of video topics. Besides that, the meta-sets have 70,10, and 21 classes for training, validation, and test sets respectively. Both different categories and the number of classes in each meta-set are promising. However, the UCF101 dataset doesn't include any additional contexts for the videos. Thus it's not suitable for the

zero-shot learning setting, which shortly aims to link different modalities. On the other hand, Figure 3.1 shows the class distributions for each meta-set in the UCF101 dataset. According to this visualization, it is clearly seen that some train and test classes are semantically close, which means the class name vectors are so close to each other in the Euclidean distance metric and highly concentrated. Even if this case may occur in real-world conditions, the overall distribution is expected to be separated at least, which can not be obtained in this dataset.

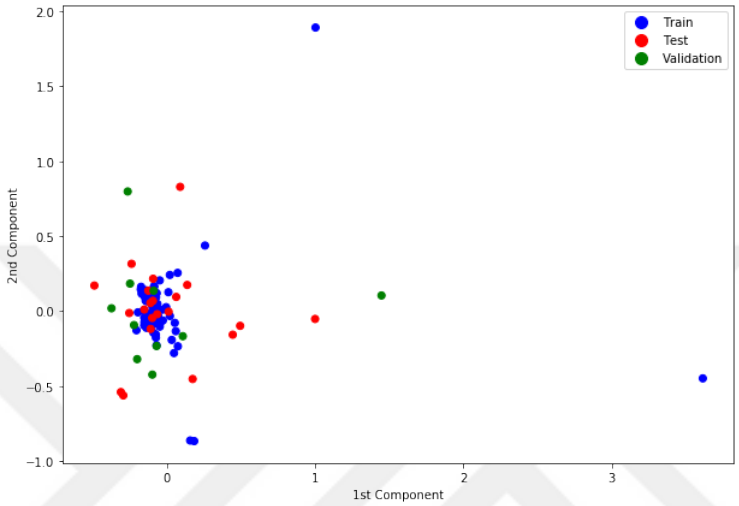


Figure 3.1 A visualization of meta-sets in UCF101 dataset.

Another dataset is HDMB-51. In this dataset, the meta-set split is done by 31,10, and 10 classes for training, validation, and test sets respectively. It has more than 100 videos for each class but the classes are oriented to human actions and their semantic similarity visualization is presented in Figure 3.2. On the additional context side, this dataset also does not provide any additional context to describe or express the action both itself and the action in the video. But actually, this dataset has meta labels for each video. This meta-label file includes 5 different properties. These properties are visible body parts, camera motion, camera viewpoint, number of people involved in the action, and video quality, and unfortunately not sufficient to identify the action itself. Because of this condition, HMDB-51 cannot be considered a dataset for ZSL problems.

Yet another and obviously large-scale dataset is Kinetics-400. This dataset consists of common and highly specified human actions with more than 300,000 video clips. This dataset was proposed for a traditional video classification problem but adapted to FSL and ZSL problems by [70]. In this adaptation process, the study randomly selected 100 classes from all possible options while ensuring that each class had 100 videos. The overall class distributions for each meta-set are presented in Figure 3.3. Similar to the previous ones, this dataset does not include the additional context.

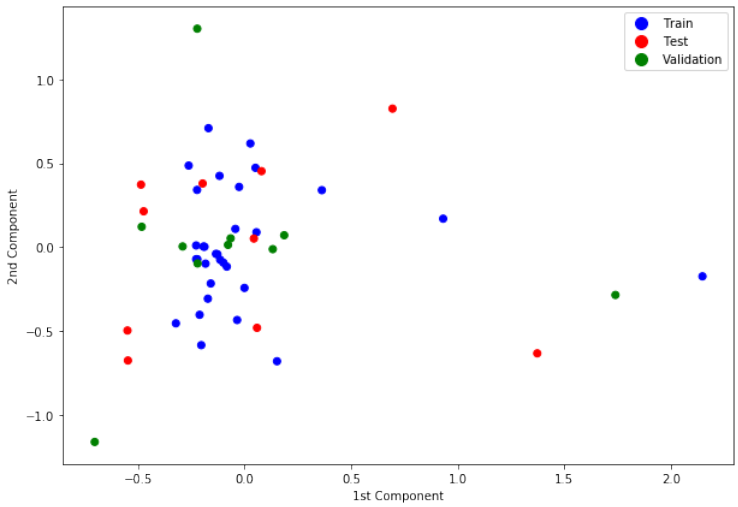


Figure 3.2 A visualization of meta-sets in HMDB-51 dataset.

Due to this reason, even if the dataset has variations and high-quality videos, this dataset can not be applied to the ZSL problem.

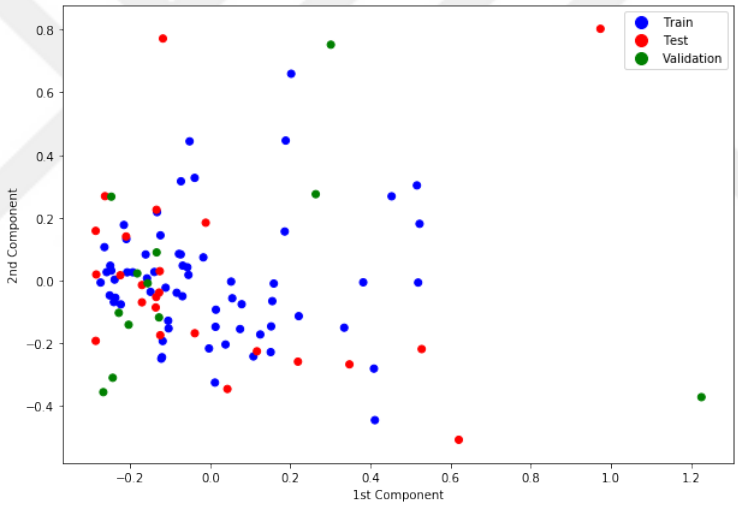


Figure 3.3 A visualization of meta-sets in Kinetics-400 dataset.

Another dataset in this research field is Something-2-Something v2. Since the dataset was constructed by a large number of crowd workers and includes pre-defined actions like *Pushing something from right to left*, has great potential for low-data-oriented problems. The dataset includes 174 different label templates that can be filled with different objects. For example, for a "Holding [something] next to [something]" template, the dataset includes different videos like "holding potato next to vicks vaporub bottle" or "holding a pencil next to tissue box", etc. And, this annotation makes it compatible and suitable for FSL and ZSL problems. The class distributions for different meta-sets of this dataset are presented in Figure 3.4. The overall structure of this dataset can be identified as suitable for FSL and ZSL problems but it is required to be improved to more distributed and varied classes.

Especially the dataset focuses the fine-grained human hand gestures-based actions. This makes this dataset less universal in FSL and ZSL datasets manner.

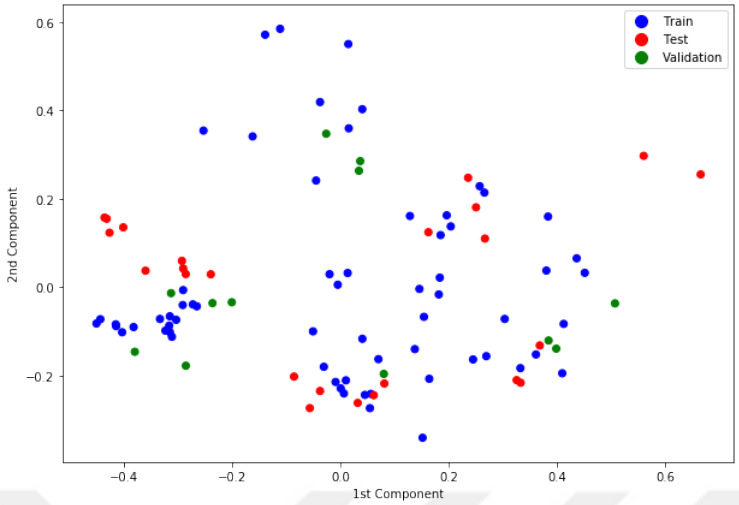


Figure 3.4 A visualization of meta-sets in Something-2-Someting v2 dataset.

In addition to all these analyses, we performed another cross-test between the training, validation, and testing subsets of the meta-sets in each data set. This cross-test was designed to find an answer to this question: "Find the average distance between the most similar 5 samples in each cross meta-sets?". More precisely, for a given each dataset, we first construct the semantic distance matrix between samples from different meta sets e.g. training and validation sets based distance matrix. Then, we searched for the indices of the 5 samples with the lowest values in the distance matrix to find the 5 most similar samples, since most similar samples have to have the lowest distance value. Then those values are averaged and presented in Table 3.1. According to the results, it is shown that our proposed dataset, which will be presented in the next section, has higher values for each meta-set. For example, in the Train-Test cross-test scenario, our proposed dataset has 0.2801, which means that most similar 5 classes have 0.2801 unit distance while this value is 0.2225 in the other datasets. Besides the Train-Test condition, our newly proposed dataset, HVU-LD, has the best scores among other previously proposed and actively used datasets. And this property of the dataset besides the hand-made analysis, demonstrates that variation within the samples across meta-sets has been improved with the new dataset. Then this ensures that the future works will be forced to get better learning abilities but this will lead the community to mimic real-world conditions better.

Following both hand-made and analytical tests, the need for a new dataset was demonstrated. In the next subsection, it will be explained in detail how this dataset was created, what pre-processing steps were applied, and what algorithm was

Dataset Name	Train - Validation	Train - Test	Validation - Test
UCF-101	0.2027	0.2225	0.3105
HMDB-51	0.2295	0.2024	0.2813
Kinetics-400	0.2065	0.1585	0.2922
Something-2-Something	0.2574	0.1723	0.2447
HVU-LD (Our Proposal)	**0.3407**	**0.2801**	**0.4754**

Table 3.1 Cross-test distance of each meta-set based on the 5 most similar samples. The scores are between 0.0 and 1.0 and represent the average distance of the 5 most similar samples in cross meta-sets. (Higher is better)

chosen during the selection of classes.

3.2 HVU-LD Dataset Construction

In the previous section, all previously proposed datasets for FSL and ZSL problems were analyzed in depth, their suitability and what they could offer for this low-data-oriented domain were investigated with different analytical methods. According to those results, we propose a new dataset, whose name is Holistic Video Understanding (HVU-LD), that provides a high variety in the class distribution, high-intra class, and low inter-class similarity-based meta-sets and mimics the real-world conditions. This dataset, HVU-LD, is constructed from the HVU dataset [26] by using several semantic methods to close the drawbacks of the previously proposed dataset in the low-data-oriented domain.

In this context, this section will focus on two key aspects. The first one is the importance of including a variety of classes in each meta-set is discussed. The second one is the process of selecting the appropriate classes for each meta-set is described. The HVU-LD dataset is a crucial resource for the low-data-based video classification domain, and understanding its construction is essential for interpreting and utilizing its results.

The first rule in FSL and ZSL settings stipulates that the classes in the meta-training, meta-validation, and meta-testing sets have to be disjoint. To satisfy this requirement, one approach is to randomly select a set of unique classes from all possible classes, ensuring that the classes in the meta-training, meta-validation, and meta-testing sets are disjoint. As stated in the [70], 100 classes for meta-training, meta-validation, and meta-testing were selected from Kinetics [51] dataset randomly. However, this randomness can cause some classes in different meta-sets to be close to each other semantically. Although semantically similar classes do not violate the disjointness rule, they do not help to build challenging

Algorithm 1 Meta-set Class Selection Algorithm

CLASS_SELECTOR($classNames$)

$classFeatures \leftarrow$ *empty list*

Extract semantic feature from each className using Fasttext method

for all $className \in classNames$ **do**
$processedClassName \leftarrow Pre_processing(className)$
$classFeature \leftarrow Fasttext(processedClassName)$
$classFeatures$.*add*($classFeature$)
end for

Apply K-means over *classFeatures* when K is 3 then obtain 3 cluster centers
$C1, C2, C3 \xleftarrow{\textit{K-Means}} classFeatures$

For each meta-set, random classes are sampled from different clusters
Meta-Train Classes $\xleftarrow{\textbf{select randomly}} C1$
Meta-Val Classes $\xleftarrow{\textbf{select randomly}} C2$
Meta-Test Classes $\xleftarrow{\textbf{select randomly}} C3$

return *[Train, Val, and Test Classes]*

real-world situations and may compromise the goal of learning completely novel classes for low-data-based learning. For this reason, we have developed the meta-set class selection algorithm, presented in Algorithm 1. The algorithm is designed to prevent the intersection of classes across different meta-sets and maximize the semantic distance between classes in separate meta-sets.

Our proposed meta-set class selection algorithm consists of two main steps, the first one is preprocessing of the raw class names and the second is semantic feature extraction.

3.2.1 Class-name Preprocessing

The textual pre-processing step includes lowercase conversion, stopword removing, alpha-numeric char correction, and stemming operation sequentially. The reason of doing such an operation is to ensure that each class name has its purest version of itself and prevent the misunderstanding of semantic features. For example, a stopword like word "the", may appear in two different class names and this may make the outputs of two class names closer to each other. Or, any punctuations that exist in two different class names may have the same effect. To get rid of these

kinds of noise situations, we first cleaned the class names with those pre-processing methods. Some examples according to this pre-processing step are presented in Table 4.1.

3.2.2 Semantic Feature Extraction

The second main step in meta-set class selection algorithm is semantic feature extraction from processed class names. Semantic feature extraction is the process of expressing the textual processed class names given as input in a vector format after some operations. The constraint here is that, as in the language databases created according to the semantic groups of this vector, the distances of word vectors that contain semantic closeness are expected to be close, while the distances of word vectors that do not are expected to be far. For example, the vectors of the words "king" and "queen" should be closer than the vector of the word "Istanbul". In this step, we preferred to use the official pre-trained English Fasttext [14] study without doing any fine-tuning or weight updating. Fasttext is a learnable word embedding study that is trained on a large corpus with the CBOW technique to capture morphological information. Also, it uses the 5-gram length that indicates the number of words considered before and after the target word in the training context. Finally, 10 negatives refer to the number of negative samples used in the training process, which helps the model learn to distinguish between true context words and random noise.

Then according to extracted semantic features, the K-Means clustering algorithm is applied. Since the aim is building a semantically distant three meta-sets, we set K number as three. This clustering operation helps to identify which classes are close to each other and which are not. Then, according to the clustering result, the classes to be selected for each meta-set are taken from different clusters. Thus, we first verify that the meta-training set has 64, the meta-validation set has 12, and the meta-test set has 24 video classes; here, there are 100 videos for each class, regardless of the set. Second, each meta-set has unique video classes, and the selected classes are as far away as possible in terms of semantic distance, which is highly desirable in a low-data-based learning problem. Ultimately, both the classes in each meta-set are chosen differently from each other, and the classes that are as semantically distant as possible are chosen among the meta-sets.

As a final step, we validated our clustering-based meta-set class selection algorithm across other previously proposed datasets to show its performance analytically. To do this, we configured a test setup that is based on the intra-class and inter-class distances of meta-sets. Distances were calculated by comparing the semantic

Table 3.2 Distance analysis of the meta-training, meta-validation, and meta-testing classes in terms of Euclidean distance.

Dataset	Intra-Class Distance	Inter-Class Distance	Overall Ratio
UCF-101	1.284	0.2977	4.313
HMDB-51 [83]	1.796	0.564	3.184
Kinetics-100 [70]	1.144	1.145	0.999
Sm2Sm-100 [70]	0.729	0.732	0.995
HVU-LD (Ours Random)	1.305	1.316	**0.991**
HVU-LD (Ours)	1.208	1.263	**0.956**

features obtained from the classes under each meta-set with each other with the Euclidean metric and taking the average of their sums. In these circumstances, we analyzed the semantic distances between classes in the meta-sets of both our proposed dataset and previously proposed ones, and we compared the results obtained using our meta-set class selection algorithm, (Algorithm 1) with those obtained by randomly shuffling classes and constructing meta-sets without using the algorithm. It shouldn't be forgotten that class names were applied pre-processing functions, which were defined in earlier this section before semantic feature extraction to provide consistency and robustness. We repeated this process 10 times and presented the scores in Table 3.2. In addition, the intra-class and inter-class distances, in Euclidean distance metric, of other existing data sets [49, 50] were also calculated, even if they are not suitable for the field of low data-based multi-modal video classification. Our final findings indicate that our dataset exhibits a better inter-class and intra-class distance ratio compared to the existing dataset. Since, while the inter-class distance increases, the semantic similarity of the meta-sets decreases, which is a desirable setting in low-data-based learning to mimic real-world conditions.

3.3 Dataset Summary

In this section, the datasets that were previously proposed, and our newly proposed dataset for FSL and ZSL problems, were taken into consideration and topics were organized in two steps.

Firstly, previously proposed datasets were analyzed. This analysis was completed by including the baseline and low-data-based versions of the relevant data sets. At these stages, the qualitative and quantitative properties of the classes in the meta-sets and their suitability for low data-based problems were tested with different tests. Comparisons between each other and their pros and cons are stated.

Secondly, as a result of the analysis carried out in the first step, it was decided to propose a new dataset. This proposed dataset is intended to fill the gaps in previous qualitative and quantitative-based tests. In this direction, 100 different classes from the HVU dataset were selected and presented as a new dataset for FSL and ZSL problems in the light of an unsupervised semantic distance-based class selection algorithm, which was completely developed by us.

The final classes for each meta-set are presented in the Appendix section Table 7.1. The complete information like class names, meta-set split, and video names for each class are presented in our GitHub repo, which is publicly available [84].

4
A SEMI-SUPERVISED VARIATIONAL AUTOENCODER FOR FEW-SHOT AND ZERO-SHOT LEARNING SETTINGS

4.1 Motivation and Problem Definition

As discussed in the literature review and dataset sections, video domain FSL and ZSL studies appear to either use low-diversity datasets for the FSL problem or completely inappropriate datasets, which don't include additional context than visual context, for the ZSL problem.

In this context, we first decide on a baseline study for our newly proposed dataset. This baseline study was adopted from an image domain study, [27], with some modifications to make that study video domain compatible. Then, a semi-supervision module for the variational autoencoder was proposed and combined with the baseline study to create our final proposed model, a Semi-supervised cross- and distribution alignment variational autoencoder (SS-CADA-VAE).

In the last part, both *n-way k-shot* based and dimension reduction outputted visual evaluations will be presented for baseline and SS-CADA-VAE models. Again, in the same section, the results of both the baseline and the improved model, SS-CADA-VAE, will be evaluated on previously proposed datasets to compare our method's performance with other methodologies.

After these evaluations, our motivation consists of two parts. The first part is to provide a basis for future FSL and ZSL studies, and the other part is to ensure consistency between future developments by expanding the use of our proposed dataset in these studies.

4.2 Proposed Method

In this study, a low-data-oriented video classification system that uses variational inference with semi-supervision regularization was proposed. This system accepts visual and textual modalities and supports both few-shot and zero-shot settings-based classification problems. The main purpose of the system is based on the metric-learning strategy with *n-way k-shot* episodic batches. The proposed system architecture is presented in Figure 4.1.

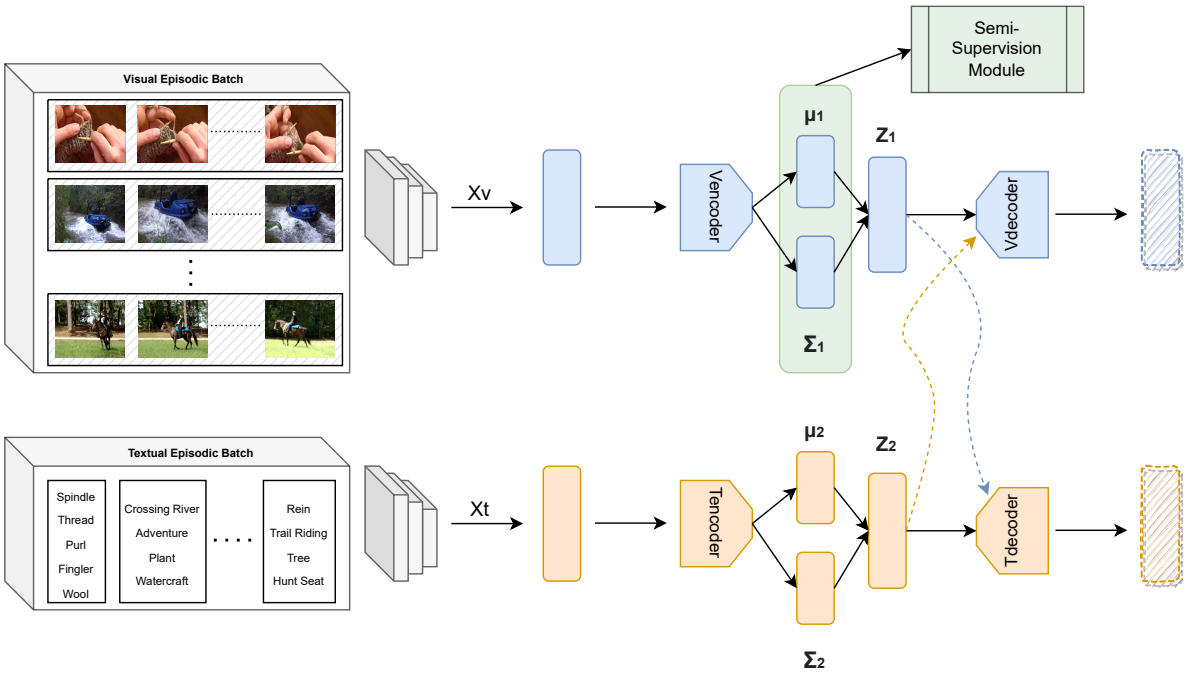


Figure 4.1 SS-CADA-VAE Model with Semi-Supervision module (SSM). SSM includes two regularization terms to enhance visual distribution parameters in terms of decreasing intra-class distance and increasing inter-class distance. In addition to regularization, SSM includes a distribution similarity-based classification operation which is used in both the few-shot setting and the zero-shot setting.

Briefly, the system consists of the following sub-systems;

- Frame selection and visual feature extraction
- Textual information preprocessing and textual feature extraction
- Variational inference & autoencoder
- Cross and distribution alignment
- NMC: Non-Parametric Metric-Learning Classifier Module

4.2.1 Frame Selection and Visual Feature Extraction

Videos are made up of frames, which can be represented as arrays with fixed dimensions. Videos differ in terms of the frames they contain, which differ in dimensions, number of depths, and depth representation. While many commercial devices vary too much in size, the vast majority use 8-bit 3-channel frames.

In this context, our proposed HVU-LD dataset consists of 8-bit 3-channel videos that have an average length of 2.1 minutes. This average length implies that there are expected 3.6k frames in each video. Processing all these frames is expensive and can be considered noisy as successive frames do not contain significant changes. Due to this reason, we selected a total of 16 frames from each video, with uniform sampling.

The next step after frame sampling is feature extraction. In this step, two different network architectures with frozen and pre-trained were preferred to measure their impacts over FSL and ZSL video classification problems. From this perspective, one convolutional neural network based, Densenet201 [85], and the transformer-based ViT [86] studies were chosen. While the Densenet201 has around 20.1M parameters, ViT has around 85.8M.

The output of the last average pooling layer for the Densenet201 study and the classification token's output for the ViT were considered as feature vectors for different architectures. Since the networks are taken as pre-trained and frozen, their proper pre-processing functions were applied.

The last point is that the feature vector that will represent the video is created by averaging the feature vectors obtained separately from each of the frames, regardless of the visual feature extraction method.

4.2.2 Textual Information Pre-processing and Textual Feature Extraction

Each video in the meta-sets includes some textual information to describe the video from different aspects. These textual modality inputs are responsible for identifying the scene, the object, etc. These inputs are used in the ZSL problem to classify videos without visual examples. In this context, we first applied certain types of pre-processing functions to normalize the textual information then used a pre-trained word embedding study to extract a feature vector for each token in the textual information.

Textual pre-processing involves cleaning and transforming raw tokens to a structured format, enhancing the performance of statistical learning models by

removing noise and standardizing inputs. This step is crucial for accurate and efficient text analysis, ensuring the data is consistent and meaningful for subsequent processing. Based on this context, the pre-processing functions, which were applied in the meta-set class selection process, like lowercasing, stopword removing, alpha-numeric character removing, and stemming are applied to normalize the raw input textual information. The example of this normalization over some inputs is presented in Table 4.1.

Table 4.1 An example of normalization over some textual inputs. The cumulative transformation of each sample input text is presented in each line

Normalization Function	Indoor_Games_ and_Sports	Flooring	Strength_Training
Lowercasing	Indoor_Games_ and_Sports	flooring	strength_training
Stopword Removing	indoor_games _sports	flooring	strength_training
Alpha-numeric Correction	indoor games sports	flooring	strength training
Stemming	indoor game sport	floor	strength train

After the textual information normalization step, two different textual feature extraction methods for video descriptions entail converting descriptive text into numerical vectors that encapsulate the semantic meaning of the text. In this thesis, besides a pre-trained FastText Continuous Bag of Words (CBOW) model, a pre-trained deep bidirectional transformer for language understanding, BERT, are employed to generate word embeddings for textual tokens of the video descriptions. These methods leverage the models' extensive training on large corpora to produce embeddings that capture both syntactic and semantic properties of the words. The use of pre-trained FastText or BERT embeddings ensures that the feature extraction process benefits from rich contextual information, thereby improving the accuracy and relevance of low-data-oriented video classification problems.

4.2.3 Variational Inference & Autoencoder

Variational inference aims to approximate intractable true posterior distribution, $p_\phi(z|x)$, using a parameterized tractable proxy posterior distribution, $q_\theta(z|x)$, while trying to maximize evidence lower bound (ELBO), or in other words, minimize the Kullback–Leibler divergence [87]. In particular, the posterior distribution of a latent variable model is obtained by the Bayes rule, which is given in Equation (4.1):

$$p_\phi(z|x) = \frac{p_\phi(x|z) \cdot p_\phi(z)}{p_\phi(x)} \ , \tag{4.1}$$

where z is a latent variable, and under these circumstances, the evidence term, $p_\phi(x)$, is intractable, and variational inference aims to approximate the posterior with another distribution in Equation (4.2):

$$p_\phi(z|x) \approx q_\theta(z|x) \ , \tag{4.2}$$

Within the scope of this equation, the approximation can be formulated as an optimization problem where θ refers to learnable parameters, and the objective function, which is a variational bound on the marginal likelihood, is presented in Equation (4.3):

$$\textit{ELBO} = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p_\theta(z)) \ , \tag{4.3}$$

The objective function in Equation (4.3) consists of two terms. The first one refers to reconstruction loss that is based on the latent variable and the second one refers to *KL* divergence between a learnable parameter-based approximated distribution and prior distribution. In our proposed method, a multivariate normal Gaussian distribution is used as a prior distribution.

4.2.4 Cross and Distribution Alignment

In a classical variational autoencoder network, there are three main components: encoder, decoder, and reparameterization module. However, in the case of multi-modal low-data-based problems, the main objective is to learn representations from multiple modalities and project them into a jointly shared space. Therefore, to achieve this, the standard variational autoencoder network is extended for each modality, and an additional mechanism is used to ensure that the output of each modality network is close to each other.

In the context of multi-modal low-data-based problems, a joint representation can be learned by training separate encoder and decoder networks for each modality. Specifically, the networks for the visual and textual modalities are denoted as Enc_{visual}, $Enc_{textual}$, Dec_{visual}, and $Dec_{textual}$, respectively. These networks enable the creation of separate latent representations for each modality, which are h_{visual} and $h_{textual}$. By training these networks to be as close as possible to

each other in the latent space, the goal is to learn a common representation that captures the underlying relationships between the modalities. This approach allows for the effective fusion of information from multiple modalities, which can lead to improved performance. To achieve this, the cross-alignment loss is proposed by [27], where the latent representation of each modality is decoded by the other modality's decoder. This approach allows for the creation of a joint representation that can better capture the underlying relationships between the visual and textual modalities.

Here, the cross-alignment loss is defined as follows:

$$\begin{aligned} \mathcal{L}_{CA_1} &= \sum_{i}^{N} \mid x_{visual}^{(i)} - Dec_{visual}(h_{textual}) \mid \\ \mathcal{L}_{CA_2} &= \sum_{i}^{N} \mid x_{textual}^{(i)} - Dec_{textual}(h_{visual}) \mid \\ \mathcal{L}_{CA} &= \mathcal{L}_{CA_1} + \mathcal{L}_{CA_2} \ , \end{aligned} \tag{4.4}$$

where $\mathcal{L}_{CA}$ aims to minimize the reconstruction loss over the (N) mini-batches when latent variables of each modality are decoded with the other modalities' decoders.

On the other hand, even though the prior distributions of each modality network are standard multivariate Gaussian distributions, the distributions of visual and textual networks may differ and, more importantly, since the two modalities' latent variables need to be projected onto the same joint shared space, approximated posterior distributions of the two modalities' networks need to be as close as possible in 2-Wasserstein distance [88]. The Wasserstein distance-based distribution alignment loss can be given as:

$$\mathcal{L}_{DA} = \sum_{i}^{N} (|| \ \mu_{visual} - \mu_{textual} \ ||_2^2 \ + \ || \ \Sigma_{visual}^{1/2} - \Sigma_{textual}^{1/2} \ ||_{Frobenius}^2)^{1/2} \ , \tag{4.5}$$

where $\mathcal{L}_{DA}$ aims to minimize the distance between the approximated posterior distribution of the visual network and the approximated posterior distribution of the textual network, over the (N) mini-batches. Equation (4.5) uses μ and Σ parameters, which are estimated from Enc_{visual} and $Enc_{textual}$.

4.2.5 NMC: Non-Parametric Metric-Learning Classifier Module

Metric-learning-based low-data learning algorithms [33, 34, 64] typically learn a feature set from both support set samples ($f_{support}$) and query set samples (f_{query}), and then compare $f_{support}$ and f_{query} by element-wise comparison using predefined distance functions such as Euclidean distance or Cosine distance. However, neither the class-based representations that are aggregated from $f_{support}$, nor the f_{query} representations that are estimated by Enc networks, take the distribution of the classes into consideration. This can lead to overfitting problems or introduce weak class-based representations that harm the classification performance.

To overcome these limitations, we propose the Non-Parametric Class Module (NMC), which estimates the class distribution for each class from $f_{support}$ instead of relying on point estimation. Specifically, NMC uses two single-layer variational autoencoder networks for visual and textual modalities and estimates a distribution for each class from the support samples of the classes. It then performs classification by comparing the query set samples with these estimated class distributions.

By using class distribution instead of point estimation-based strategies [33, 34, 64], our proposed approach creates class representations that are less susceptible to overfitting and are better suited to handle low-data scenarios. This results in improved classification performance, as demonstrated by our experimental results.

4.2.5.1 Class-Based Distribution Estimation

In metric-learning-based algorithms, the primary goal is to learn a function *F(.)* that takes the input data which can be an image or attribute in any form, then extracts a representation regarding the input data. Then, based on the representations extracted from both the support and query set samples, class representations are built using an aggregation technique, e.g. averaging, on the representations of the support set samples. In the classification stage, the class representations and query set samples' representations are compared under some distance metrics, like Euclidean, Cosine, etc. Query set samples are classified at the end using distance values or distance values-based probabilities.

From the aforementioned perspective of the existing metric-learning-based algorithms, we propose a distribution-based class representation instead of element-wise averaging of support set sample representations. In our work, Enc_{visual} and $Enc_{textual}$ represent the input data x_{visual} and $x_{textual}$ in [μ_{visual}, σ^2_{visual}] and [$\mu_{textual}$, $\sigma^2_{textual}$] forms, respectively. Using the posterior distribution of a single sample, we estimate the aggregated class distribution parameters from

support set samples as follows:

$$\mu = (\frac{1}{N}) \sum_{i}^{N} (\mu^{(i)}),\ \sigma^2 = \sum_{i}^{N} ((\frac{1}{N^2}) \cdot \sigma^{(i)^2}) \ , \tag{4.6}$$

According to Equation (4.6), class distribution parameters are estimated from support set samples for each class in each episode. More precisely, there are *n* classes in the support set, where each has *k* samples. Then from the *k* samples for the class *j* in the episode, that class distribution is as follows. $\mathcal{N}|(\mu_j, \sigma_j^2)$ is expanded from a single class *j* = [1,*k*] to all *n* classes in the current episode.

Furthermore, apart from estimating the class distribution from the support set samples, each sample in the query set is transformed into [μ, σ^2] form by utilizing either Enc_{visual} or $Enc_{textual}$. To prevent any bias resulting from a low number of tests, we used 15 samples per class in the query set for the *n*-way *k*-shot setting.

4.2.5.2 DDC: Distribution Distance-Based Classification

After estimating the class distributions as described in Section 4.2.5.1, the query set samples are classified using a distribution similarity approach. Each query sample can be represented as $D_{query} = \{(\mu_{visual_{ij}}, \sigma^2_{visual_{ij}}, \mu_{textual_{ij}}, \sigma^2_{textual_{ij}}) : | : i \in [1..15], ; j \in [1..n]\}$. Then, according to the scenario, the KL divergence is calculated between the aggregated class distributions and the distributions of the corresponding query set samples. Then, the probability that each query sample belongs to a class is determined by performing logarithmic normalization with the softmax function of those deviation scores.

The general form of KL divergence between two continuous random variables is as follows:

$$\mathcal{D}_{\mathcal{KL}}(p||q) = \int_x p(x) \cdot \log(\frac{p(x)}{q(x)}) \ . \tag{4.7}$$

Then, according to the probability density function of multivariate normal distribution, the finalized version of *KL* divergence is as follows:

$$\mathcal{D}_{\mathcal{KL}}(p||q) = \frac{1}{2}[\log(\frac{|\Sigma_q|}{|\Sigma_q|}) - 1 + \ \ (\mu_p - \mu_q)^T \Sigma_q^- 1 \, (\mu_p - \mu_q) + tr\Sigma_q^- 1 \Sigma_p] \ , \tag{4.8}$$

And the probability calculation between a query set sample *i* and the class *m* in the

n-way *k*-shot setting is calculated using Equation (4.8) as follows:

$$Pr_\theta(y = m \mid \mathcal{N}_i(\mu_i, \sigma_i^2)) = \frac{\exp(\mathcal{D}_{\mathcal{KL}}(\mathcal{N}_i||\mathcal{N}_m))}{\exp(\sum_{j=1}^{n} \mathcal{D}_{\mathcal{KL}}(\mathcal{N}_i||\mathcal{N}_j))} \ , \tag{4.9}$$

In this context, the normalized probability of a query set sample over *n* classes is calculated. Then, the classification label is calculated as follows:

$$\hat{y} = \arg\max_{m \in n} Pr_\theta(y = m|\mathcal{N}_i(\mu_i, \sigma_i^2)) \ , \tag{4.10}$$

4.2.6 SS-CADA-VAE Loss

The semi-supervised cross and distribution-aligned VAEs consist of multiple objectives to enable low-data-based video classification. The overall objective is as follows:

$$\mathcal{L}_{SS-CADA-VAE} = \mathcal{L}_{VAE} + \alpha\mathcal{L}_{CA} + \beta\mathcal{L}_{DA} + \gamma\mathcal{L}_{SS} \ , \tag{4.11}$$

In Equation (4.11), α, β, and γ refer the weights of the corresponding objective functions.

4.2.7 Implementation Details

In this thesis, CADA-VAE was chosen as a baseline for low-data video classification. Originally designed for the still-image domain, some adaptations were necessary for its video application. CADA-VAE consists of two encoders and two decoders for each modality, each with one hidden layer. For video input, frames are extracted and visual feature vectors are computed using different feature extractors. These vectors are then averaged to create a video-level feature vector, allowing CADA-VAE to be used in the video domain with minimal changes to its architecture.

In order to analyze the effect of different visual backbones on the low-data-based video classification problem, two different neural network architectures have been used to obtain videos for frames' visual-level features: Densenet201 [85] and visual transformer [86]. These architectures are pre-trained on ImageNet. And they are completely frozen during the low-data learning process. Regardless of the feature extraction methodology, 16 frames were selected uniformly for each video in uniform order and then resized to 224 $\times$ 224. Then, to obtain the video-level

visual feature vector, each of the selected frames is fed into the feature extraction method and their output is averaged. Similarly, regardless of the feature extraction method, the visual hidden layer has 960 neurons in the case of Densenet201 and 384 neurons in the case of ViT usage. Hereby, while the network that uses the Densenet201 feature extraction method has nearly 4 M learning model parameters, the network that uses ViT has nearly 3 M learning model parameters.

At the same time, with the same intention, two different pre-trained neural network architectures have been used to obtain videos for frames' textual-level features: ELMO [89] and Bert [90]. While in ELMO, the videos' textual-level feature is constructed by averaging each token's ELMO features, we only use the classification head feature vector in the Bert method. Furthermore, without depending on the textual neural network architecture, the textual hidden layer has 512 neurons to analyze videos' textual-level features. And, the latent hidden size for both modalities was set to 64. The result of the number of learning parameters of the textual network has nearly 1.5 M learning model parameters due to both ELMO and Bert models being frozen during the low-data learning process.

On the other hand, L2 loss is chosen as a reconstruction and a cross-reconstruction objective. While the reconstruction objective's coefficient is set to 1.0 during the training, the coefficient of the cross-reconstruction, α, increases up to 1.5 between the 20th and 75th epochs. And, the coefficient of KL-divergence in the ELBO objective is set to 0.25, while the coefficient of the distribution alignment objective, β, increases up to 3 between the 10th and 30th epochs. Also, the semi-supervision part of the SS-CADA-VAE consists of the γ coefficient. Different combinations and thresholds have been examined and detailed experiments are presented for both the few-shot and the zero-shot classification tasks. Since using multiple loss objectives simultaneously in any learning process may cause gradient conflict and damage training stability. In this context several different approaches have been proposed to adjust this usage in deep neural networks [91, 92]. To prevent this behavior and maximize the benefit of different loss objectives we first analyzed the learning curves, which belong to testing and validation classification losses, in each epoch. Besides these curves, the gradient flow of the loss functions are also observed in each step to detect any anomalies. All these observation processes are handled with Tensorflow Tensorboard tool, which is presented in this study's Github repo. In light of this information, the overall coefficients of the loss objectives are presented in Figure 4.2.

All components in the SS-CADA-VAE are trained for 100 epochs by the first-order derivative-based Adam optimizer, while the learning rate is 1×10^{-3}. Then,

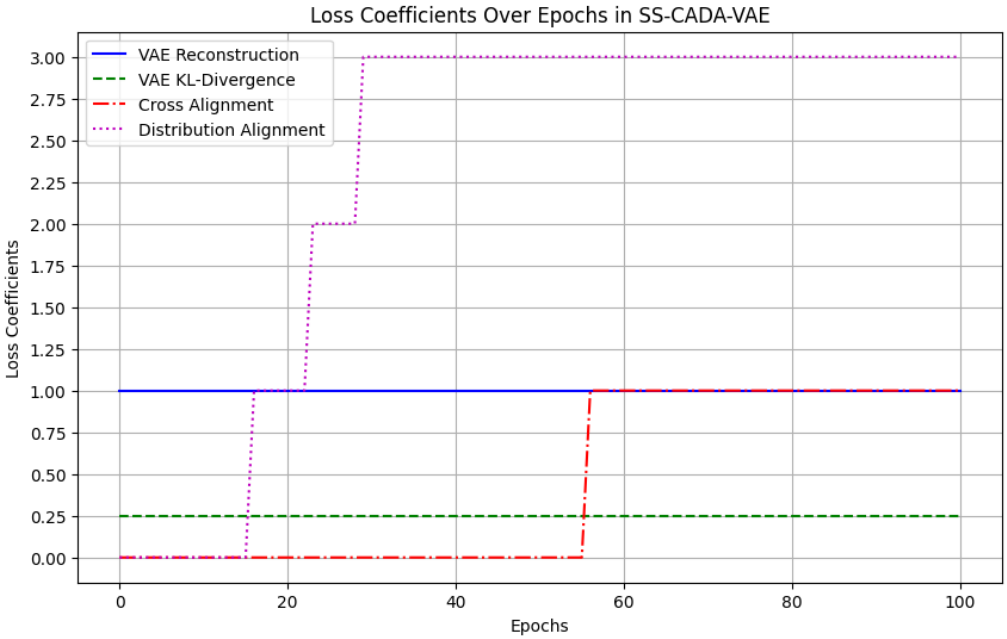


Figure 4.2 The coefficients of the loss objectives in SS-CADA-VAE are presented with their initial and finalized values, and increment strategy (Semi-Supervision objective aka $\mathcal{L}_{SS}$, has been experimented with different values, which are presented in the experiments section).

episodic learning is selected throughout the entire model. In this manner, each episode has n classes, where each class has k samples in the support set. These parameters depend on the learning scenario, for example, while *3-way 5-shot* is one of the few-shot learning options, *3-way 0-shot* is the zero-shot learning option. In each learning scenario and setting, 15 samples are preferred in the query set, since a low number of test samples may produce biased observations. Also, during the testing stage, we randomly construct 300 episodes for one trial and repeat this trial 10 times to obtain confident test scores.

5
EXPERIMENTS

In this section, we first define the evaluation pipeline for our proposed study and dataset. Even if our study has a similar term with traditional supervised learning models, the classification term, the evaluation procedure is completely different from the evaluation procedure of traditional supervised learning models. In traditional supervised learning models, the models are expected to generalize the internal patterns of the training data and are then tested on data from the same distribution as the training data, but never used in the training phase. The final performance of the model consists of the metric results obtained from that group of data. But we use *n-way k-shot* episode-based evaluation procedure for low-data-oriented problems. The first and most critical part of this procedure is that the training set is completely different from the validation and test sets in both distribution and class. And in the testing phase, each episode contains n different classes and k examples for each class. The classification process is carried out based on the k samples presented for each n class, with m examples presented for each class that is completely different from these k classes. For the sake of example, let the video classification problem be *3-way 5-shot*. In this condition, there are 3 different classes, 5 samples for each class, and m different testing samples for each class. The episodic evaluation expects to classify m samples for each class correctly by taking the 5 samples for each class. The final remarkable point between FSL and ZSL is, in the FSL video classification problem, while the k is bigger than zero in each episode, k is set to zero in the ZSL video classification problem. Then it means that ZSL expects to classify the videos by not considering the visual frames, but only using their textual information. The pseudo-explanation of the FSL and ZSL test stages is presented in Algorithm 2. Also, the overall system architectures for the FSL and ZSL are presented in Figure 5.1.

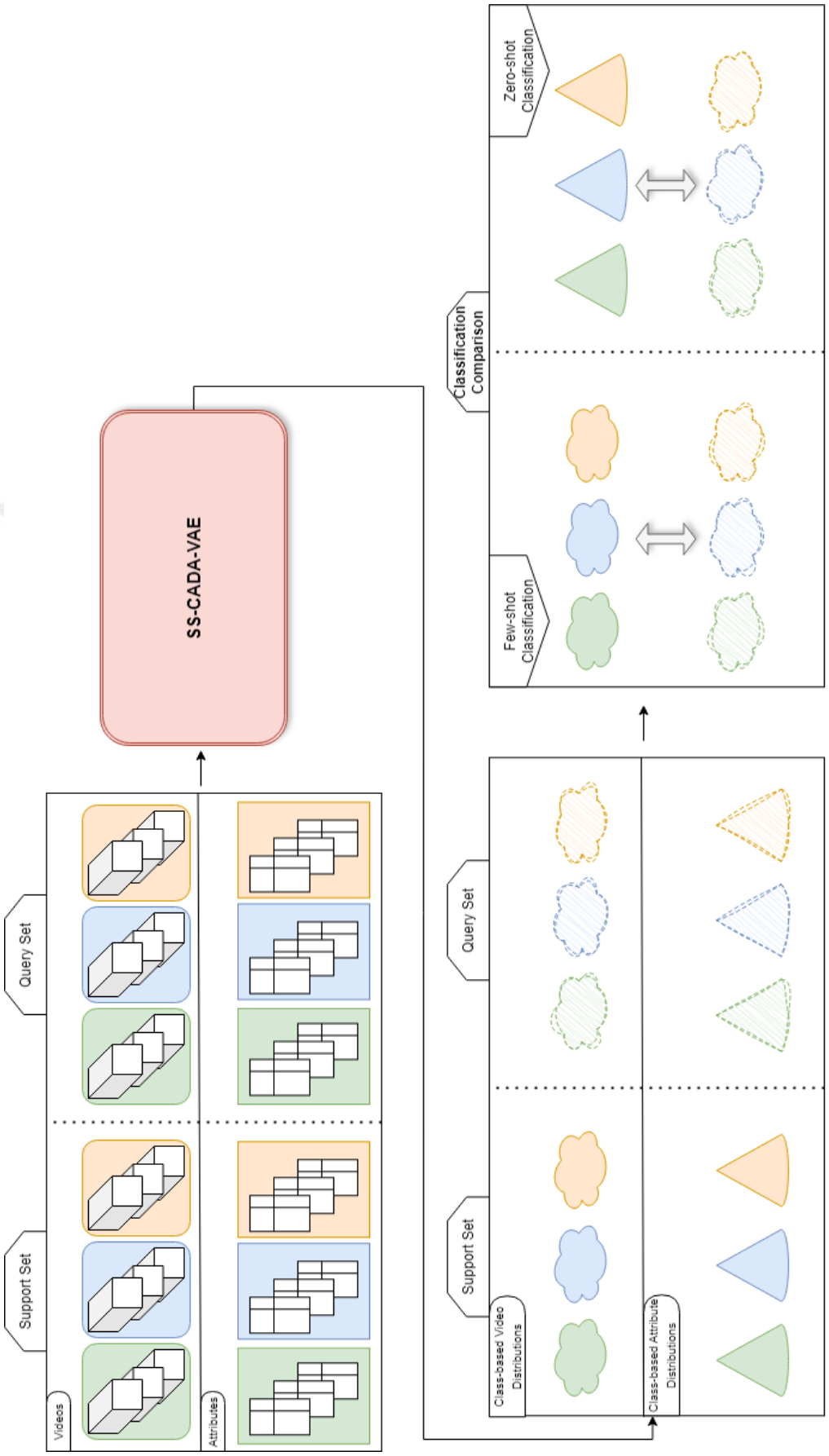


Figure 5.1 The system architecture of SS-CADA-VAE in the FSL and ZSL test classification stages. Both stages share the common distribution construction step, but then they differ in the context of the modalities used during classification.

Within the light of this pipeline overview, we first evaluate the SS-CADA-VAE approach on the newly proposed HVU-LD dataset, which includes additional modalities beyond class labels not present in existing video classification datasets, we conducted experiments in both few-shot learning and zero-learning settings for 0-shot, 1-shot, 3-shot, 4-shot, 5-shot, and 10-shot scenarios with varying numbers of support set samples, as well as in 3-way, 5-way, and 10-way scenarios. In the evaluation stage, episodes are selected randomly. This may cause a biased situation for some certain classes. To alleviate this problem and propose a more robust testing score we repeat each test 10 times and report the average accuracy as the final result.

Also, we compare several different SS-CADA-VAE hyper-parameters to perform a comprehensive ablation study, testing various combinations of α, β, and γ parameters to analyze their impact on low-data-based video classification. We set the baseline model as a CADA-VAE [27] with the same hyper-parameters as the original model, excluding the γ parameter, which is the semi-supervision enhancement module in the SS-CADA-VAE. Additionally, we use the *Distribution Distance-Based Classification* (DDC) method, as shown in Section 4.2.5.2, to compare the improvements made by the SS-CADA-VAE more sharply. Besides all those combinations, we also compare the visual and textual feature extraction methods to investigate their impacts on low-data-based video classification.

At the same time, we mentioned the importance of dataset class selection for different meta-sets in the dataset section. Based on this point, we also kept a log of the accuracy performance of meta-set classes. In this way, it was also investigated whether class choices affected performance.

On the other hand, we also evaluated our SS-CADA-VAE model with the previously existing uni-modal dataset in the few-shot learning scenario to verify whether our proposed model is competitive with state-of-the-art few-shot video classification models. Since our model is proposed to provide a baseline for the newly proposed low-data-based multimodal dataset, it should have competitive results with state-of-the-art models, which are proposed for uni-modal-oriented low-data-based studies, so that it can be characterized as a solid starting point for future studies.

5.1 Few-Shot Classification Results

In this section, FSL results will be provided with different hyper-parameters; number of classes in the episode, *n-way*, and number of samples for each class, *k-shot*. The general trade-off between these two hyper-parameters can be summarized

Algorithm 2 Episodic Testing for Few-Shot and Zero-Shot Learning

function EpisodicTesting(model, test_data, N, K, attribute_data=None)
    $episodes \leftarrow$ GenerateEpisodes$(test_data, N, K)$
    $accuracy_list \leftarrow []$
    for each $episode \in episodes$ **do**
        $support_set, query_set \leftarrow episode.support, episode.query$
        $support_labels \leftarrow \{example.label \mid example \in support_set\}$
        if FSL **then** ▷ Few-Shot Learning
            $support_embeddings \leftarrow model.embed(support_set)$
        else ▷ Zero-Shot Learning
            $support_attributes \leftarrow$ GetAttributes$(support_set, attribute_data)$
        end if
        $correct_predictions \leftarrow 0$
        for each $query \in query_set$ **do**
            $query_embedding \leftarrow model.embed(query.input)$
            if FSL **then** ▷ Few-Shot Learning
                $prediction \leftarrow$ model.predict(
                    $query_embedding,$
                    $support_embeddings,$
                    $support_labels$
)
            else ▷ Zero-Shot Learning
                $prediction \leftarrow$ model.predict(
                    $query_embedding,$
                    $support_attributes,$
                    $support_labels$
)
            end if
            $correct_predictions \overset{\text{append}}{\leftarrow}$ is_correct$(prediction, query.label)$
        end for
        $accuracy \leftarrow \frac{correct_predictions}{\text{len}(query_set)}$
        $accuracy_list.append(accuracy)$
    end for
    return mean$(accuracy_list)$
end function

as that when the number of classes increases, the classification problem becomes challenging due to find out the discriminative features, and the number of samples for each class directly affects the classification performance like when it increases the classification performance becomes easier or vice versa.

Besides these hyper-parameters, the visual and the textual feature extraction backbones play a crucial role in the FSL problem. The general intuition behind feature extraction backbones relates to their ability to discriminate between different input samples.

During this experiment, we tested both the low-data-oriented video classification system hyper-parameters and the feature extraction backbones and presented their effects on the FSL problem. Also, our proposal, DDC module 4.2.5.2, which is a non-parametric distribution distance-based classification module for FSL and ZSL problems, was tested with all these previously expressed hyper-parameters. During this module's experiments, we also took the [27] model as a baseline.

5.1.1 Impact of Visual Feature Methodologies and Distribution Distance-Based Classification Module for FSL

In this analysis section, two different hyper-parameters will be investigated. The first and the main subjective is the visual-level feature extraction methods and their impact on the FSL problem. As indicated and deeply analyzed before, Densenet201 and ViT are used for both FSL and ZSL problems in this thesis. The second subjective of this analysis section is our proposal, the distribution distance-based classification module. For that purpose, the $\mathcal{L}_{SS}$ coefficient in Equation 4.11, which is γ coefficient, will have two values; 0 and 1. When this coefficient is set to 0, it means that the distribution distance-based classification module is disabled, in other words, the baseline model, while 1 refers to the opposite.

Also, the whole proposed neural network structure 4.1 has 2 main modalities, which are visual and textual. To make robust tests, the textual-level features are set as fixed with ELMO features. Thus, we tried to eliminate the effect of the textual-level feature extraction operation on visual-level feature extraction methods.

In this concept, the results for *5-way* few-shot learning are presented in Table 5.1. In that table, whereas the models that use Densenet201 as a visual backbone have *CNN* tag, the models that use visual transformer have *ViT* tag. Also, there are 2 more different tags; *BL* and *I*. While BL refers to the baseline model, which doesn't use the distribution distance-based classification module, I-BL refers to the model that uses that proposed module.

Table 5.1 Results of *5-way* few-shot video classification on the meta-test set. (Results are presented in percentage).

Model Type	1-Shot	3-Shot	5-Shot	10-Shot
BL CNN	51.42	71.38	77.25	82.88
I-BL CNN	57.28	76.23	81.22	85.03
BL ViT	65.59	82.30	85.97	88.51
I-BL ViT	71.05	84.89	87.81	89.85

According to the results, we can conclude visual backbone and distribution distance-based classification module-based conclusions. For all *k-shot* tests, it's observed that the visual transformer-based model has better scores than Densenet201. Especially, for the lower k-shot settings, which is the hardest condition, the visual transformer-based model has more than 14% accurate. On the other side, our proposal, the distribution distance-based classification module, has positive effects for both Desnet201 and ViT-based models. On average, *I-BL* models, which use the distribution distance-based classification module, have nearly 5% performance boost for each model structure. Besides these observations, there is another situation that requires attention is the effect of *k-shot* over the visual feature extraction backbone and the distribution distance-based classification module. It is seen that when the number of samples for a class increases from 1 to 10, the effect of those things decreases dramatically. This behavior is highly expected and understandable cause when the number of samples increases, the problem becomes more like a classical classification problem instead of the low-data-oriented classification problem. Then, the improvements become less valuable, and data dominance emerges.

5.1.2 Impact of the Number of Classes in the Episode and Distribution Enhancement Module for FSL

In this analysis section, the number of courses in a department will be investigated. In low data-driven problems, this number is highly related to the complexity of the problem and starts from 3 and ends at 10 in the literature. Therefore, we set this number to 3, 5, and 10 separately and repeated the tests.

Besides that, we also repeated the same test protocol for visual and textual feature extraction backbones and the distribution distance-based classification module. In this way, the test will show the different hyper-parameters and their effect for different numbers of classes.

There is another point to consider at this point, which is the number of samples

for each class. This number, *k-shot* can be nameable as a size training set like in a traditional classification problem. And, when this number increases the problem becomes a traditional classification problem, especially for k » 10. For this reason and in order to fully measure the effect of different hyper-parameters, we set this *k-shot* number to 1 in all tests. Thanks to this value, which is the most difficult situation in the classification problem with little data, we were able to measure the performance of the study under the most challenging conditions by obtaining the minimum performance value. The result of this analysis is presented in Table 5.2.

Table 5.2 Results of *1-shot* few-shot video classification on the meta-test set for different *n-way* settings. (Results are presented in percentage).

n-Way	BL CNN	I-BL CNN	BL ViT	I-BL ViT
3-way	63.91	69.26	76.43	79.43
5-way	51.42	57.28	65.59	71.05
10-way	37.68	43.11	52.38	58.91

As shown in Table 5.2, $\mathcal{L}_{SS}$ regularization, in other names the effect of the distribution distance-based classification module, improves both the Densnet201-based few-shot learning classification problem by an average of 5.55% across all *n-way* settings and the ViT-based few-shot learning classification problem by an average of 4.99% across all *n-way* settings in the 1-shot setting. Additionally, when $\mathcal{L}_{SS}$ regularization-free scores are examined, it is seen that, just like the result of the previous test, CNN-based results lag behind ViT-based results.

5.1.3 Impact of the DDC Module's Coefficient for FSL

One another analysis is related to the DDC, the distribution distance-based classification module. In this study, we proposed to use the DDC module in baseline study [27] to enhance the class distributions. With this improvement, even if the number of samples is low like 1-shot or 3-shot, the neural network structure aims to represent each class in an episode, the same classes closer to each other in the distribution space, and different classes farther from each other.

In order to perform such a test, we repeated the protocol in previous tests and measured the performance gains at different values of the coefficient of $\mathcal{L}_{SS}$, γ, in the error equation given in Equation 4.11. The results for different γ values are presented in Table 5.3.

The $\mathcal{L}_{SS}$ regularization coefficient starts from 0.0 and goes up to 1.0 with a step size of 0.25. During the testing, the *5-way* classification problem was implemented with

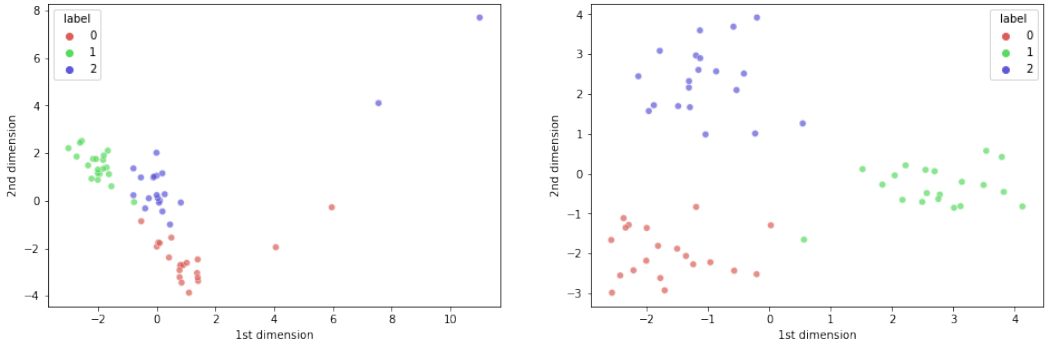
Table 5.3 Results of *5-way* few-shot video classification on the meta-test set for different γ coefficient values on CNN-based architecture with different k-shot values. (Results are presented in percentage).

	1 Shot	3 Shot	5 Shot	10 Shot	Average
γ = 0.0 (BL)	51.42	71.38	77.25	82.88	70.73
γ = 0.25	60.91	78.54	82.68	85.77	76.97
γ = 0.50	63.70	77.96	82.10	84.80	77.14
γ = 0.75	47.44	66.39	72.20	78.03	66.01
γ = 1.0	57.28	76.23	81.22	85.03	74.94

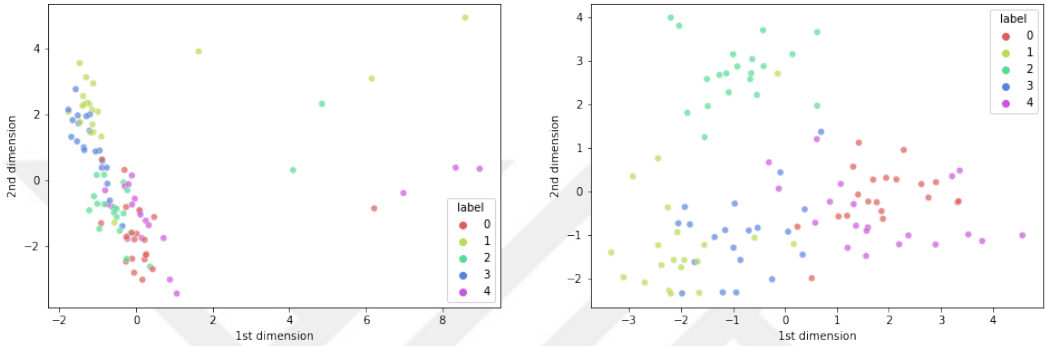
1, 3, 5, and 10 shots only with the CNN-based visual feature extractor, Densenet201. In this test scenario, the γ value of 0.0 actually refers entirely to the baseline study, with the video-domain adaptations, while different states of this value show the effect of the module we propose.

According to the classification scores in Table 5.3 the average scores for each shot value of the baseline study is 70.73%. But when the coefficient is 0.5, the classification performance increases up to 6.41%. Even if the improvement effect of this coefficient does not follow a linear trend, that is, the classification performance increases in the same direction with each increase of the variable, it is proven that the module we propose at the points 0.5 and 0.25 provides a visible performance contribution to the FSL problem.

Besides this analytic testing, we also experimented a visual testing of the effect of our proposed regularization module. In this manner, we randomly select the episodes and encode those videos with trained and frozen visual encoder Enc_{visual}, Densenet201. Then using the linear discriminant analysis algorithm, we reduced the sizes of latent vectors.



(**a**) A sample from the HVU-LD meta-testing set for *3-way 5-shot* setting.



(**b**) A sample from the HVU-LD meta-testing set for *5-way 5-shot* setting.

Figure 5.2 Visual analysis of the impact of the DDC module on the HVU-LD meta-test

The top-left and bottom-left sub-figures in Figure 5.2 show how the samples will be distributed in the 2D reduced plain by using only the baseline model or, in other words, without using the distribution distance-based classification module *DDC*. The top-right and bottom-right figures are 2D projections of the samples that are used in the left figures using the distribution distance-based classification module *DDC*. It's been observed that the samples encoded with the DDC module have better inter-class distance and intra-class cohesion.

5.2 Zero-Shot Classification Results

In this analysis section, the results for the zero-shot, in other means *0-shot*, classification will be presented. In the zero-shot video classification scenario, the classes in the support set contain only textual information, unlike the few-shot scenario. However, the query set contains only visual information. At this point, the zero-shot video classification scenario aims to learn a distribution from the textual information in the support set and match it with a completely different modality, the visual distribution, in the query set.

During this analysis, both visual feature extraction backbones, $\mathcal{L}_{SS}$ regularization, and different textual feature extraction backbones will be examined. Same as the previous evaluation protocol, *BL* models refer to the γ equals 0, and *I-BL* models have the DDC module with γ equals 0.5, which is the best value according to the tests in Table 5.3. According to this definition, the results of the zero-shot video classification is presented in Table 5.4

Table 5.4 Results of zero-shot video classification on the meta-test set for different visual and textual architectures (results are presented in percentage).

n-Way	BL CNN	I-BL CNN	BL ViT	I-BL ViT
ELMO	32.46	41.06	37.03	43.42
Bert	**39.72**	**42.35**	**48.91**0	**52.19**

According to the results, it's seen that the performances for both textual-level feature extraction backbones show a similar trend to their contextualized power in other downstream tasks [93]. However, when the $\mathcal{L}SS$ regularization objective is applied, we observe a significant improvement in zero-shot classification performance. Specifically, the use of $\mathcal{L}SS$ leads to an increase of 8.6% in CNN and ELMO-based architecture, 6.39% in ViT and ELMO-based architecture, 2.63% in CNN and Bert-based architecture, and 3.28% in ViT and Bert-based architecture, respectively.

In addition, when comparing the performance of ELMO and Bert models independently of the DDC module, Bert is more than 7% successful in the CNN model and more than 11% successful when used with ViT. In addition, it can be seen that with the use of the DDC module, the CNN and ELMO-based architecture can exceed the basic results of the CNN and Bert-based architecture. In this way, in addition to increasing the performance, the DDC module also provides efficiency in terms of calculation by increasing the performance of the model with lower parameters.

5.3 Few-Shot Classification on Existing Methods and Datasets

In the previous analysis sub-sections following topics have been analyzed;

- Our newly proposed low-data-oriented video classification dataset, HVU-LD
- Baseline study and our newly proposed distribution distance-based module onto the baseline study

- Convolutional neural network and visual transformer-based visual feature extraction backbones and their effects on FSL and ZSL video classification problems
- Deep neural network and masked language model oriented transformer architecture-based textual feature extraction backbones and their effects on FSL and ZSL video classification problems

In addition to the fact that the dataset we proposed supports FSL and ZSL video classification problems, our contribution to the baseline study, and these comprehensive tests with different hyper-parameters, we decided to test the performance of the newly proposed model with previously proposed datasets.

At this stage, as mentioned in the previous sections, the previously proposed datasets are unfortunately not suitable for the zero-shot learning problem. The main reason is that they do not contain data in any other modality than visual modality that can identify the videos. For this reason, we will compare our proposed model and related improvements with the latest work in this field within the scope of the few-shot video classification problem.

In this context, the generalization ability of the proposed model is important and should be able to achieve a competitive result, at least when the uni-modal low-data-based video classification, which is the closest research subject to the proposed dataset, is compared to the most successful models in the research area.

In this manner, we have evaluated the BL model that does not have our proposed regularization technique and is adapted to the video domain from [27], and the I-BL baseline, which has the regularization technique, with a very well-known dataset in that research area. In order to make this evaluation happen, we first obtain the raw videos of the HMDB-51 dataset. Then, according to [83], meta-train, meta-validation, and meta-validation sets are arranged. Then, as presented in our proposed dataset construction stages, ViT-based visual features for the meta-set videos are extracted. Finally, the classification results of both the BL and our I-BL models and other studies in uni-modal few-shot video classification studies are presented in Table 5.5.

The results of the HMDB-51 dataset, which are presented in Table 5.5, have similar patterns in comparison between *BL* and *I-BL* in terms of classification performance with our newly proposed dataset, HVU-LD. More precisely, while the BL model under-performs for all *k-shot* classification settings, with the help of the regularization technique, the I-BL model has better scores between nearly

10% and 17%. Besides the BL and I-BL comparison, our proposed method, I-BL, has 2.6% better classification performance in a 5-way 5-shot classification setting. Besides that, our model has gained 3.77% better classification performance among presented 5-way 3-shot classification results.

Table 5.5 Results of *5-way* few-shot video classification on the HMDB-51 meta-test set on ViT-based architecture with different k-shot values. (Results are presented in percentage.) "-" indicates that there is no value presented in the original paper for that experiment.

	1 Shot	3 Shot	5 Shot	10 Shot
OTAM [94]	54.5	65.7	68.0	-
HYRSM [83]	60.3	71.7	76.0	-
STRM [95]	-	-	81.30	-
FSVC-ATA [96]	**65.78**	-	82.27	-
BL	42.92	64.47	71.55	79.02
I-BL(ours)	60.01	**75.47**	**84.87**	**89.14**

5.4 FSL and ZSL Classification Final Remarks

In this final analysis section, we will take a step back and look at the results of both FSL and ZSL video classification problems from a distance. The purpose of conducting such a study is to detect some errors that cannot be obtained from all previous analytical experiments, the relationships between these errors, or the effect of the relationship between the videos / textual information on the performance.

In this context, we first start with the FSL video classification problem. In order to get the hidden relations behind the videos, textual information, and their effects on the performance, we used the best Densenet201 and ELMO-based architecture that uses the distribution distance-based classification with the best γ value. And, we preferred *10-way 5-shot* video classification setting. The reason we prefer this evaluation setting is to obtain robust information by comparing as many classes as possible at the same time. Besides that, the test is repeated 300 times, and for each test, 150 test samples, in other means videos and their textual information, were used. In total, 45000 episodes where each episode has 10 classes with 5 training samples and 15 testing samples, were used to perform this FSL video classification test. The confusion matrix of this evaluation is presented in Figure 5.3.

According to the confusion matrix, the weighted F1-Score is 0.6512 and the accuracy score is 0.6460. However, while some classes are well above these average lines, some are performing worse than these lines. For example, *cleaning toilet*,

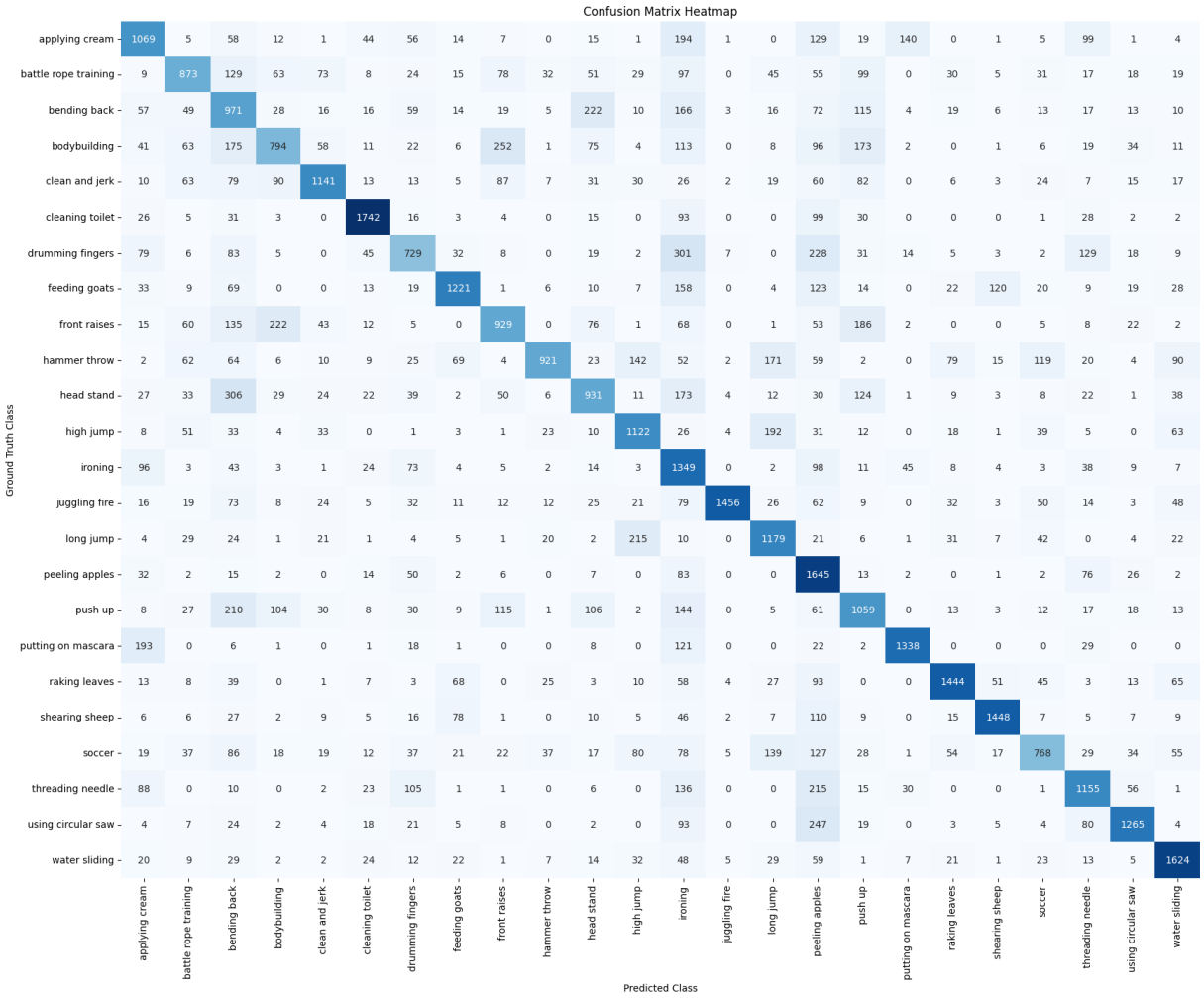


Figure 5.3 The confusion matrix of FSL for 45k episodes. (Best viewed in colored and zoomed-in.)

juggling fire, and *shearing sheep* are the most correctly classified video classes. On the other hand *bending back*, *bodybuilding*, *drumming fingers*, and *ironing* are the classes that have the worst classification performance.

Again, regarding this situation, the number of errors between some classes is interesting. For example, the error rate between *putting on mascara* and *applying cream* is striking. Also, *bending back* and *headstand* have the same striking relation. Whenever we analyze all those errors, it can be concluded that the classes that have similar actions and/or use similar objects, tend to be misclassified more frequently. A visual example of this error type is presented in Figure 5.4.

This FSL evaluation proves our ideology in the construction of the HVU-LD dataset. In the dataset proposing stage, we aimed to choose classes that are as close as possible to each other in the semantic distance to create challenging and similar to the real word conditions.



Figure 5.4 A visual example between 'bending back' and 'headstand' classes. (Best viewed in colored and zoomed-in.)

We also performed a similar evaluation for the ZSL video classification problem. In this evaluation, we followed the same number of episodes and test protocols as we did in the FSL video classification. However, unlike FSL, where class distributions were created from the support videos in the episodes, in ZSL, we compared the class distributions created from textual information instead of the videos in the query set. Specifically, we derived class descriptions from textual data and used these for the classification task, aiming to evaluate the model's ability to classify unseen classes based on their semantic descriptions. The confusion matrix for this ZSL evaluation is presented in Figure 5.5.

The confusion matrix analysis of our ZSL video classification model reveals critical insights into the model's performance when comparing video-based class distributions to textual information. Misclassifications, such as those between *Applying Cream* and *Clean and Jerk* or *Battle Rope Training* and *Push Up* suggest challenges in aligning visual features from videos with semantic information derived from textual information. These errors arise due to the inherent difficulty in mapping detailed visual movements and contexts to their corresponding textual descriptions.

For example, agricultural activities like *Feeding Goats* and *Shearing Sheep* share overlapping visual features and textual descriptors related to farm environments, contributing to misclassification. Figure 5.3 shows the visual frames of the misclassification between those classes.

Also, according to the visual example, first, we obtained the textual descriptions of two classes and then found the intersection between those textual descriptions.

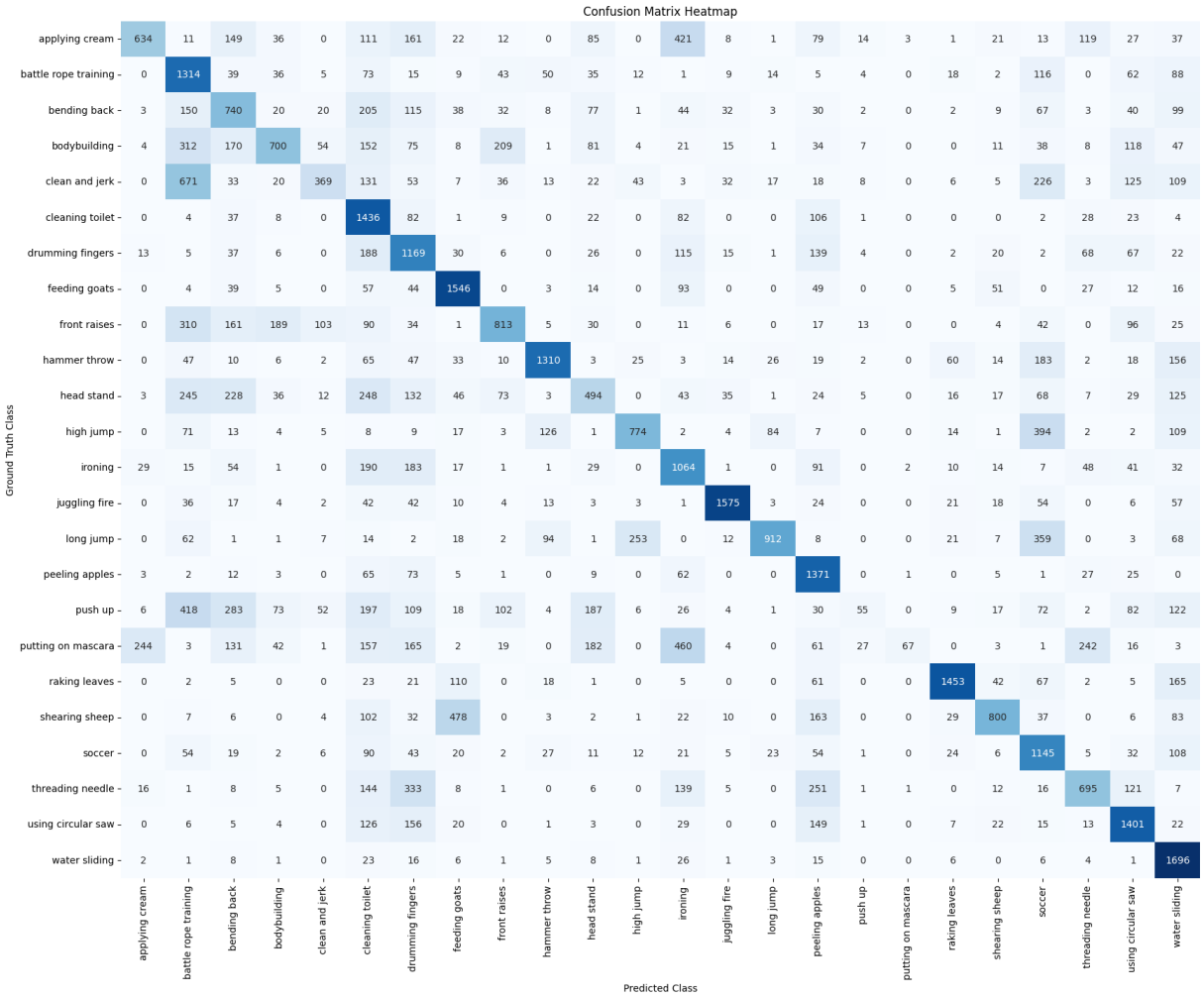


Figure 5.5 The confusion matrix of ZSL for 45k episodes. (Best viewed in colored and zoomed-in.)

Since the main learning goal of ZSL video classification is to create a semantic bridge between visual features and textual descriptions, vectorial representations of *feeding goats* and *shearing sheep* classes were created from the textual information of both their unique and intersecting elements using the ELMO model. These 3 groups were then visualized using the dimensionality reduction method, which was employed in the previous experiments. These numerical representations are given in Figure 5.7.

This analysis highlights the need for more nuanced textual descriptions and enhanced mapping techniques between visual and textual data to improve the ZSL model's accuracy. Addressing these issues is crucial for advancing the robustness and reliability of ZSL in video classification tasks. In addition, as we mentioned in the dataset creation section, current datasets are either not suitable for ZSL video classification or have tiny differences in the semantic space instead of having large distances between the classes selected for FSL video classification. For this

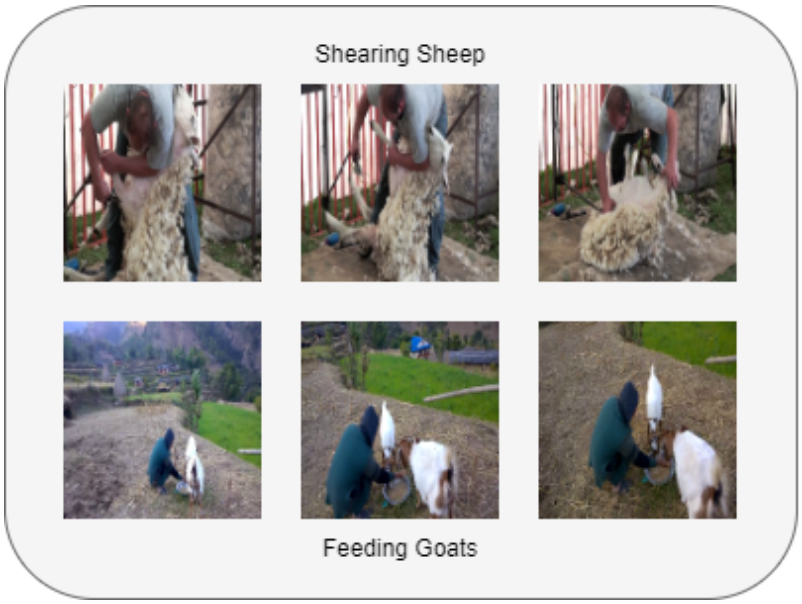


Figure 5.6 A visual example of misclassification between 'feeding goats' and 'shearing sheep' classes. (Best viewed in colored and zoomed-in.)

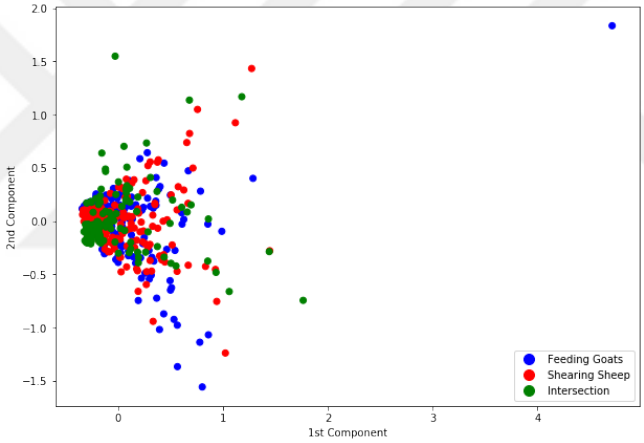


Figure 5.7 Numerical representations of the unique and intersected textual descriptions of feeding goats and shearing sheep classes.

reason, we have again demonstrated the need for a dataset that can mimic real-world conditions.

6
CONCLUSION

According to the advancement in the machine learning and computer vision research domains, it's clearly seen that low-data-oriented applications are getting too much attention. These applications can be from different areas, from classification problems to segmentation problems or object detection problems. Regardless of the type of problem, researchers always desire to achieve the results obtained by SOTA models with minimal dataset requirements. The allure of these applications lies in their potential to perform effectively even when large annotated datasets are not available, which is often the case in real-world scenarios. This shift towards low-data solutions is not only practical but also essential for democratizing advanced machine learning technologies, making them accessible and applicable in diverse fields and regions where data collection resources are limited.

The domain of video classification presents multifaceted challenges, primarily driven by the vast and intricate nature of visual data within videos. This complexity translates to substantial computational demands and necessitates extensively annotated datasets for effective classification. Traditional approaches, while effective to some extent, often fall short when faced with limited data scenarios.

In this context, this thesis aimed to propose a novel, multi-modal dataset specifically designed for low-data-based video classification tasks, including few-shot and zero-shot learning and a variational autoencoder model that is improved with the distribution distance-based regularization method.

The primary objective of proposing a new dataset was to address the limitations of existing datasets in the field, which do not adequately support both FSL and ZSL video classification problems. Furthermore, our experiments demonstrated the critical importance of class selection and its significant impact on performance when constructing datasets specifically designed for low-data-oriented problems. By carefully selecting classes for our dataset, we aimed to emulate the challenging conditions encountered in real-world scenarios, thereby ensuring that the dataset is

robust and effective for low-data video classification tasks.

The next proposal of this thesis was the distribution distance-based regularization module. With this module, the mean and variance calculation of a traditional variational autoencoder is forced to keep the intra-class distribution closer and the inter-class distribution far from each other. In this way, this module, which was added to the baseline model, provided a performance increase of up to 12% in some evaluations of FSL and ZSL protocols.

We want to create a starting point for researchers who will continue their work in this field with both the new dataset and the proposed model proposed in the thesis. We have seen that there have been studies with different evaluation protocols, with different datasets and datasets that are unfortunately not suitable for FSL and/or ZSL, and we carried out our relevant work to regulate this area.

6.1 Future Works

We see that today's computer vision studies are based on large language model studies, especially for the FSL and ZSL fields. In this field, studies in the field of natural language processing with ChatGPT [97] have been reflected in the vision-language model pair such as PeliGemma, which is a compose of a vision [98] and a language model [99], etc. and emerged with various applications in different fields.

In the initial training stages of these large models, datasets whose sizes can be expressed as TBs are generally used. Later, they can find a place for themselves in different fields with different tuning methods and smaller data set expectations.

At this point, based on the tests we have conducted, we think that you can reach higher performance points with a modality-agnostic feature extractor based on transformer models. In addition, the ability to tune and use these large models, which were just mentioned, after making modifications that can analyze videos, appears as another possible option that can increase performance. However, it should not be forgotten that the data used during the training of these large models should be carefully analyzed just before the tuning process, so that a data leak does not occur by mixing it with the classes and/or videos in the test set we offer.

7
APPENDIX

Train	Validation	Test
air drumming	adjusting glasses	applying cream
alpine skiing	basketball moves	battle rope training
beach soccer	bowling	bending back
bmx	choreography	bodybuilding
boating	climbing	clean and jerk
calf roping	directing traffic	cleaning toilet
cooking egg	equitation	drumming fingers
crochet	gymnastics	feeding goats
disc dog	hoverboarding	front raises
diving cliff	making a sandwich	hammer throw
docking boat	playing basketball	head stand
downhill mountain biking	salsa dancing	high jump
eating cake		ironing
eating ice cream		juggling fire
feeding fish		long jump
fishing		peeling apples
flying kite		push up
golf		putting on mascara
golf putting		raking leaves
grooming dog		shearing sheep
grooming horse		soccer
hula hooping		threading needle
ice climbing		using circular saw
ice fishing		water sliding
luge		
making tea		
milking cow		
mountain biking		
outdoor recreation		
playing beer pong		
playing polo		
playing water polo		
pumping gas		
purl		
rafting		
riding a bike		
riding bumper cars		
riding elephant		
riding mower		
riding or walking with horse		
riding scooter		
riding unicycle		
roasting pig		
rock climbing		
roller skating		
scuba diving		
shoveling snow		
skateboarding		
skiing cross-country		
skiing mono		
skiing slal om		
snow tubing		
tai chi		
tasting beer		
tasting wine		
training dog		
visiting the zoo		
wading through water		
walking the dog		
walking through snow		
watching tv		
water skiing		
windsurfing		
zumba		

Table 7.1 Classes of the meta-sets of the HVU-LD low-data-oriented dataset

REFERENCES

[1] W. S. McCulloch, W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, pp. 115–133, 1943.

[2] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[3] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[4] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[5] J. Long, E. Shelhamer, T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[6] F. Liu, C. Shen, G. Lin, I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2024–2039, 2015.

[7] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[8] Y. Yu, H. Ko, J. Choi, G. Kim, "End-to-end concept word detection for video captioning, retrieval, and question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3165–3173.

[9] J. De Fauw *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[10] H. Proença, J. C. Neves, "Irina: Iris recognition (even) in inaccurately segmented data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 538–547.

[11] D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[12] G. Hinton, *Rmsprop*, `http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf`, Accessed: 2021-05-17, 2015.

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

[14] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[15] J. P. Chiu, E. Nichols, "Named entity recognition with bidirectional lstm-cnns," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 357–370, 2016.

[16] X. Glorot, A. Bordes, Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *ICML*, 2011.

[17] S. Hochreiter, J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] A. Vaswani *et al.*, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[19] G. Badaro, M. Saeed, P. Papotti, "Transformers for tabular data representation: A survey of models and applications," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 227–249, 2023.

[20] K. Zaman, M. Sah, C. Direkoglu, M. Unoki, "A survey of audio classification using deep learning," *IEEE Access*, 2023.

[21] H. Luo *et al.*, "Univl: A unified video and language pre-training model for multimodal understanding and generation," *arXiv preprint arXiv:2002.06353*, 2020.

[22] C. Lyu *et al.*, "Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration," *arXiv preprint arXiv:2306.09093*, 2023.

[23] E. Alpaydın, *Introduction to machine learning second edition*, 2010.

[24] S. Abu-El-Haija *et al.*, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[25] S. Thrun, L. Pratt, "Learning to learn: Introduction and overview," in *Learning to learn*, Springer, 1998, pp. 3–17.

[26] A. Diba *et al.*, "Large scale holistic video understanding," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 593–610.

[27] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, Z. Akata, "Generalized zero-and few-shot learning via aligned variational autoencoders," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8247–8255.

[28] *MS Windows NT kernel description*, `https://www.mturk.com/`, Accessed: 2024-01-01.

[29] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[30] B. M. Lake, R. Salakhutdinov, J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[31] H. Zhang, J. Zhang, P. Koniusz, "Few-shot learning via saliency-guided hallucination of samples," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2770–2779.

[32] A. Li, T. Luo, T. Xiang, W. Huang, L. Wang, "Few-shot learning with global class representations," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9715–9724.

[33] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, *et al.*, "Matching networks for one shot learning," in *Conference on Neural Information Processing Systems*, 2016.

[34] J. Snell, K. Swersky, R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 4080–4090.

[35] F. Pahde, M. Puscas, T. Klein, M. Nabi, "Multimodal prototypical networks for few-shot learning," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2021, pp. 2644–2653.

[36] Y. Zhang, H. Tang, K. Jia, "Fine-grained visual categorization using meta-learning optimization with sample selection of auxiliary data," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 233–248.

[37] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, L. Van Gool, "One-shot video object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 221–230.

[38] C. Finn, P. Abbeel, S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*, PMLR, 2017, pp. 1126–1135.

[39] B. Lake, R. Salakhutdinov, J. Gross, J. Tenenbaum, "One shot learning of simple visual concepts," in *Proceedings of the annual meeting of the cognitive science society*, vol. 33, 2011.

[40] S. Ravi, H. Larochelle, "Optimization as a model for few-shot learning," in *International conference on learning representations*, 2016.

[41] P. Welinder *et al.*, "Caltech-ucsd birds 200," 2010.

[42] G. Patterson, J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 2012, pp. 2751–2758.

[43] C. H. Lampert, H. Nickisch, S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *2009 IEEE conference on computer vision and pattern recognition*, IEEE, 2009, pp. 951–958.

[44] Y. Xian, C. H. Lampert, B. Schiele, Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2251–2265, 2018.

[45] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[46] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[47] E. Dataset, "Novel datasets for fine-grained image categorization," in *First Workshop on Fine Grained Visual Categorization, CVPR. Citeseer. Citeseer*, Citeseer, 2011.

[48] J. Krause, M. Stark, J. Deng, L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.

[49] K. Soomro, A. R. Zamir, M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv 2012, arXiv:1212.0402*,

[50] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, "Hmdb: A large video database for human motion recognition," in *Proceedings of the International Conference on Computer Vision*, 2011, pp. 2556–2563.

[51] W. Kay *et al.*, "The kinetics human action video dataset," *arXiv 2017, arXiv:1705.06950*,

[52] C. H. Fabian, E. Victor, G. Bernard, C. N. Juan, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 961–970.

[53] R. Goyal *et al.*, "The "something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5842–5850.

[54] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, J. Sivic, "HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips," in *ICCV*, 2019.

[55] *Wikihow*, `https://www.wikihow.com/`, Accessed: 2024-01-01.

[56] *Google cloud vision api. available online: `\NoCaseChange{https://cloud.google.com/vision}`*, (accessed on 1 February 2024).

[57] *Sensifai video tagging api. available online: `\NoCaseChange{https://sensifai.com}`*, (accessed on 1 February 2024).

[58] H. Zhang, J. Zhang, P. Koniusz, "Few-shot learning via saliency-guided hallucination of samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2770–2779.

[59] A. Li, T. Luo, T. Xiang, W. Huang, L. Wang, "Few-shot learning with global class representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9715–9724.

[60] Y.-X. Wang, R. Girshick, M. Hebert, B. Hariharan, "Low-shot learning from imaginary data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7278–7286.

[61] K. Li, Y. Zhang, K. Li, Y. Fu, "Adversarial feature hallucination networks for few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 470–13 479.

[62] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, D. Wierstra, "Matching networks for one shot learning," *arXiv preprint arXiv:1606.04080*, 2016.

[63] J. Snell, K. Swersky, R. S. Zemel, "Prototypical networks for few-shot learning," *arXiv preprint arXiv:1703.05175*, 2017.

[64] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208.

[65] W. Xu, Y. Xian, J. Wang, B. Schiele, Z. Akata, "Attribute prototype network for zero-shot learning," *arXiv preprint arXiv:2008.08290*, 2020.

[66] Y. Chen, X. Wang, Z. Liu, H. Xu, T. Darrell, "A new meta-baseline for few-shot learning," *arXiv preprint arXiv:2003.04390*, 2020.

[67] F. Pahde, M. Puscas, T. Klein, M. Nabi, "Multimodal prototypical networks for few-shot learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2644–2653.

[68] Y.-H. Hubert Tsai, L.-K. Huang, R. Salakhutdinov, "Learning robust visual-semantic embeddings," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3571–3580.

[69] S. Chen *et al.*, "Transzero: Attribute-guided transformer for zero-shot learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 330–338.

[70] L. Zhu, Y. Yang, "Compound memory networks for few-shot video classification," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 751–766.

[71] K. Cao, J. Ji, Z. Cao, C.-Y. Chang, J. C. Niebles, "Few-shot video classification via temporal alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 618–10 627.

[72] M. Muller, "Information retrieval for music and motion," 2007.

[73] R. Jin, X. Wang, G. Wang, Y. Lu, H.-M. Hu, H. Wang, "Embedding adaptation network with transformer for few-shot action recognition," in *Asian Conference on Machine Learning*, PMLR, 2023, pp. 515–530.

[74] X. Wang *et al.*, "Clip-guided prototype modulating for few-shot action recognition," *International Journal of Computer Vision*, pp. 1–14, 2023.

[75] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763.

[76] X. Wang *et al.*, "Few-shot action recognition with captioning foundation models," *arXiv 2023, arXiv:2310.10125*, 2023.

[77] J. Li, D. Li, C. Xiong, S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*, 2022, pp. 12 888–12 900.

[78] Y. Tang, B. Béjar, R. Vidal, "Semantic-aware video representation for few-shot action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 6458–6468.

[79] Y. Xian, B. Korbar, M. Douze, B. Schiele, Z. Akata, L. Torresani, "Generalized many-way few-shot video classification," in *European Conference on Computer Vision*, Springer, 2020, pp. 111–127.

[80] B. Thomee *et al.*, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[81] B. Brattoli, J. Tighe, F. Zhdanov, P. Perona, K. Chalupka, "Rethinking zero-shot video classification: End-to-end training for realistic applications," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4613–4623.

[82] C.-C. Lin, K. Lin, L. Wang, Z. Liu, L. Li, "Cross-modal representation learning for zero-shot action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 978–19 988.

[83] X. Wang *et al.*, "Hybrid relation guided set matching for few-shot action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 948–19 957.

[84] E. Citak, *Fewshotmultimodal*, `https://github.com/ErolCitak/FewShotMultiModal`, 2024.

[85] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[86] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv 2020, arXiv:2010.11929*,

[87] D. P. Kingma, M. Welling, "Auto-encoding variational bayes," *arXiv 2013, arXiv:1312.6114*,

[88] C. R. Givens, R. M. Shortt, "A class of wasserstein metrics for probability distributions," *Michigan Math. J.*, vol. 31, no. 2, pp. 231–240, 1984.

[89] M. E. Peters *et al.*, "Deep contextualized word representations," *arXiv 2018, arXiv:1802.05365*,

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

[91] C. Xu *et al.*, "Multi-loss regularized deep neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 12, pp. 2273–2283, 2015.

[92] Q. Cheng, H. Li, Q. Wu, K. N. Ngan, "Hybrid-loss supervision for deep neural network," *Neurocomputing*, vol. 388, pp. 78–89, 2020.

[93] N. Reimers, B. Schiller, T. Beck, J. Daxenberger, C. Stab, I. Gurevych, "Classification and clustering of arguments with contextualized word embeddings," *arXiv 2019, arXiv:1906.09821*,

[94] K. Cao, J. Ji, Z. Cao, C.-Y. Chang, J. C. Niebles, "Few-shot video classification via temporal alignment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 618–10 627.

[95] A. Thatipelli, S. Narayan, S. Khan, R. M. Anwer, F. S. Khan, B. Ghanem, "Spatio-temporal relation modeling for few-shot action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 958–19 967.

[96] K. D. Nguyen, Q.-H. Tran, K. Nguyen, B.-S. Hua, R. Nguyen, "Inductive and transductive few-shot video classification via appearance and temporal alignments," in *European Conference on Computer Vision*, 2022, pp. 471–487.

[97] OpenAI, *Chatgpt: GPT-4 large language model*, Accessed: 2024-06-14, 2024. [Online]. Available: `https://www.openai.com`.

[98] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 975–11 986.

[99] G. Team *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

PUBLICATIONS FROM THE THESIS

Papers

1. E. Citak, M. E. Karsligil, "Multi-Modal Low-Data-Based Learning for Video Classification," Applied Sciences, vol. 14, no. 10, pp. 4272, 2024.

2. E. Citak, M. E. Karsligil, "Few-Shot and Zero-Shot Video Classification from Large Visual Models Perspective," in Proceedings of the International Conference on Engineering, Natural Sciences, and Technological Developments, 2024, pp. 848-854.