applied sciences

MDPI

Article

Multi-Modal Low-Data-Based Learning for Video Classification

Erol Citak * and Mine Elif Karsligil

Computer Engineering Department, Yildiz Technical University, 34349 Istanbul, Turkey; elif@yildiz.edu.tr
* Correspondence: erol.citak@std.yildiz.edu.tr

Abstract: 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 categories. However, due to the vast amount of visual data contained in videos, the classification process is often computationally expensive and requires a significant amount 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 challenges. However, existing low-data area datasets, which are either not diverse or have no additional modality context, which is a mandatory requirement for the zero-shot task, do not fulfill the requirements for few-shot and 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. The dataset contains pairs of videos and attributes that capture multiple facets of the video content. Thus, the new proposed dataset will both enable the study of low-data-based video classification tasks and 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. In addition, 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.

Keywords: multi-modal dataset; few-shot learning; zero-shot learning; video classification; multi-modal learning; deep variational autoencoder

Citation: Citak, E.; Karsligil, M.E. Multi-Modal Low-Data-Based Learning for Video Classification. *Appl. Sci.* **2024**, *14*, 4272. https://doi.org/10.3390/app14104272

Academic Editor: Samuel Cheng

Received: 24 March 2024
Revised: 22 April 2024
Accepted: 25 April 2024
Published: 17 May 2024

1. Introduction

Video classification has received a lot of attention due to its promising real-world practical applications such as human activity classification [1], medical examination [2], autonomous vehicles [3], or video recommendation [4,5]. Despite outstanding results and applications, the video classification field requires a large amount of labeled data in the training stage. Even though existing semi-supervised labeling mechanisms [6] or data augmentation techniques are proposed in the literature, label acquisition is still a time-consuming and expensive stage. Due to this reason, low-data-based deep learning models have attracted a lot of attention, especially in the video classification field.

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 a few-step learning scenario, 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 context from other modalities.

According to these limitations, it is a vital requirement whether a dataset exists that will enable multi-modal low-data video classification. This dataset must have the auxiliary context to help analyze the video from different perspectives, as in the still image domain, for video classification problems based on multi-modal and zero-shot learning. In this context, the existing video classification datasets [7–9] only have class labels and do not provide auxiliary context about the videos. However, multi-modal low-data-based video classification research requires a dataset that contains both videos and additional modality context that identifies videos (e.g., natural language-based text content that semantically describes videos), which is a key contribution of this paper.

In the light of all these requirements and deficiencies in the field of research, the contributions of this study are listed below:

- First, we propose a dataset, due to the fact that current video classification datasets are highly biased towards specific actions, which provides low diversity, and lack auxiliary context beyond class labels to enable proper zero-shot learning and multi-modal learning. Our dataset is based on the Holistic Video Understanding dataset [10], 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. Training, validation, and testing stages are applied to the meta-training, meta-validation, and meta-testing sets, respectively. As shown in Figure 1, 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 they do not have 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 not only that 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 is applicable 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 [11], 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.

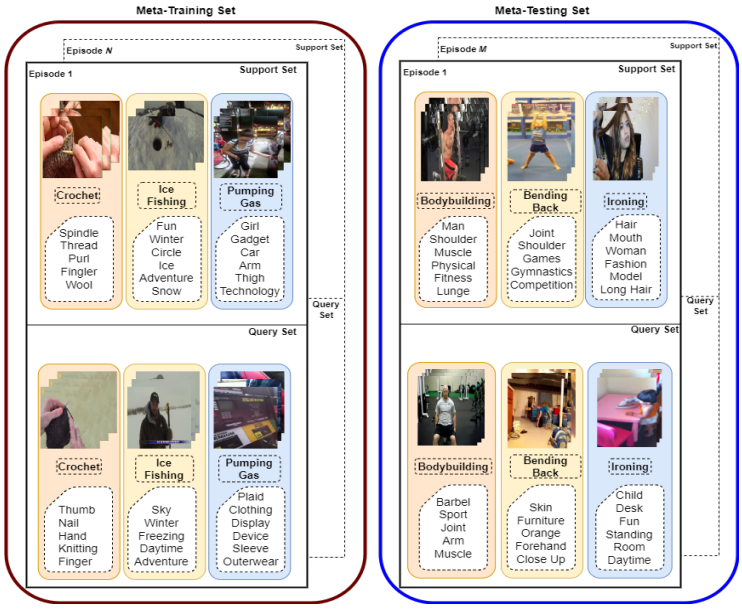


Figure 1. The meta-learning paradigm-based multi-modal low-data-based video classification setting. Meta-training and Meta-testing sets follow the *n-way k-shot* setup. The intersection of these sets is empty, and the learning model is trained with Meta-training episodes and tested with meta-testing episodes. Also, each episode both in training and testing sets has support and query sets where the support set refers to training examples, and the query set refers to testing samples. (Best view colored and zoomed in).

2. Background and Related Work

Low-data-based supervised learning models differentiate from conventional supervised learning models in terms of the number of available samples in the training stage. To address this challenge, low-data-based learning algorithms have been developed with three different primary perspectives [12]. The most intuitive strategy is data augmentation, which aims to increase the number of samples either using hand-crafted techniques like flipping, color changes, rotation changes [13], or novel learning model-based data hallucination models [14,15]. Another strategy to alleviate the low-data-oriented overfitting issue is using external data and a more sophisticated fine-tuning strategy [16,17] which tries to revisit the fine-tuning pipeline for low-data problem adaptation. Finally, a comparison-based approach known as learning to compare [18–20] has shown promising results. This approach uses comparisons to enhance learning by defining prototype representations and comparing test samples to these prototypes.

2.1. Few-Shot and Zero-Shot Learning in Still-Image Domain

In the context of few-shot learning and zero-shot learning, the sets of classes used for training and testing the model are distinct. This differs from the standard supervised learning pipeline, where the same set of classes is used for both training and testing. As a result, models used in few-shot and zero-shot learning scenarios must be able to generalize to new classes that were not seen during training. This presents a unique challenge for machine learning algorithms, as they need to learn and generalize from a limited number of examples or textual descriptions of new classes.

In the few-shot meta-learning setting, each episode consists of k samples from n distinct classes, and the learning process utilizes these samples during the training phase. One approach proposed in [19] involves combining the visual features of the given samples for each class to construct class prototype representations. These prototypes are then used

in distance metric-based classification. In addition to the uni-model approach, some works have incorporated auxiliary context during the training stage, such as in [20,21]. Auxiliary context typically provides a multi-faceted description of the still image and has been used to improve the quality of class representations or to enhance the visual representation capacity of the model. Ref. [22] uses supervised contrastive loss with the video transformer [23] to handle the few-shot image classification problem.

In zero-shot learning, unlike few-shot learning, model evaluation is operated over the cross-modalities. More precisely, in zero-shot learning, no samples are given at testing time, and the model is 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. By this shared feature space, different modalities are enforced to have joint representation for zero-shot evaluation. In this manner [24], the model generates visual and textual features with the help of autoencoders, which have two branches for each modality for classification purposes, and then it minimizes the maximum mean discrepancy between visual and textual modality feature space distributions to obtain joint latent space feature. Alongside the latent distribution alignment, ref. [11] implements a cross-reconstruction loss which follows the fact that each encoded feature for all modalities is decoded in all modality decoders. Ref. [25] implements the attribute-guided transformer mechanism for the visual modality, then in the decoding stage of this transformer, language model-based textual features are embedded into visual features to obtain visual–semantic joint latent space.

2.2. Few-Shot and Zero-Shot Learning in Video Domain

In the domain of video, ref. [26] proposed a dataset for few-shot learning to close the gap in this research area using the Kinetics dataset [9]. 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. Ref. [26] 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. While using this newly created dataset and the Something-V2 [27] dataset, Tam [28] proposed a novel long-term temporal ordering information-based comparison module with the help of the DTW algorithm [29] on top of embedded visual features of both support and query set videos. Through temporal alignment distance, ref. [28] accomplishes the classification purpose of the few-shot learning scenario. Ref. [30] follows the embedding network architecture and proposes an instance normalization to improve the discriminative power of the feature. On the other hand, ref. [31] employs the pre-trained [32] model, which is trained to connect text and images into a shared space, to extract multi-modal features. Then, the study includes prototype modulation with those features to solve the few-shot action recognition problem. Ref. [33] generates synthetic video captions with a foundation model, by ref. [34], for the datasets that have no auxiliary content, like captions. Then, in light of these video captions, they propose a multi-modal few-shot action recognition solution.

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 specific and mostly human-based actions: UCF101 [8], HMDB51 [7], ActivityNet [35]. 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, ref. [36] proposed an end-2-end architecture that uses Fasttext word embeddings [37] to extract continuous embedding from the action class label, and unlike the previous works in zero-shot video classification, it implements a 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 the 3D-CNN, ref. [38] proposes a cross-modal transformer network that is able to extract visual–semantic joint space representation simultaneously.

3. A New Multi-Modal Video Classification Dataset

The topic of few-shot learning and zero-shot learning using still images has gathered significant attention and has resulted in the creation of large-scale and high-variation benchmark datasets such as CUB [39], SUN [40], AWA1 [41], and AWA2 [42]. In the video domain, there are some datasets for conventional video classification problems, like HMDB-51 [7], UCF-101 [8], and Kinetics-400 [9]. Even if these datasets have proven to be excellent candidates for the conventional video classification problem, which does not take the low-data situation into consideration, such datasets are not suitable for video-based low-data learning since they only provide class labels of videos as auxiliary context. The deficiency caused by the absence of auxiliary information prevents low-data-based multi-modal video classification studies. Because the existence of video-oriented and detailed auxiliary context is a necessity for both multi-modal classification and zero-shot video classification problems, to address this gap, we propose the HVU-LD dataset based on the Holistic Video Understanding (HVU) dataset [10] to ensure the consistency of future learning algorithms' performances. In this regard, we analyze the HVU dataset, discuss why the HVU-LD dataset is proposed for low-data-based video classification, and provide a detailed description of the dataset construction process, including class selection and pre-processing operations applied to the auxiliary context.

3.1. HVU Dataset

The 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 572 k videos and 9 million annotations covering 3142 labels across six 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 six semantic categories, a two-stage framework was employed that involved both machine-generated automatic labeling and human-based validation processes. The Google Vision API [43] and Sensifai Video Tagging API [44] 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 1. Each row indicates the category name, the number of distinct labels per category, and the number of videos that have that category-oriented label.

Table 1. The overall distribution per category number of labels and the number of videos.

Category	# of Labels	# 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

3.2. Why Do We Need the HVU-LD 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. The task definition of ZSL is expressed as follows. Let $D_s = \{(x, y, a(y)) \mid x \in X^s, y \in Y^s, a(y) \in A^s\}$ is the meta-training set consisting of visual features x, class labels y, and auxiliary context $a(y)$, where s denotes the seen classes. At the same time, the meta-testing dataset can be defined as follows: $D_u = \{(x, y, a(y)) \mid x \in X^u, y \in Y^u, a(y) \in A^u\}$, where u denotes the unseen classes and the intersection of Y_s, and Y_u is a null set. In the inference stage, the learning model aims 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 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. An example for a meta-learning episode is given in Appendix A.

At this point, the first zero-shot learning and few-shot learning settings need a large-scale and high-variation dataset that can mimic real-world scenarios. However, existing datasets [7,8,35] in low-data-based video classification are highly biased towards human-oriented actions or specific sports actions and do not cover many other actions in real-world cases. Additionally, as presented above, the zero-shot learning setting requires rich auxiliary context, A^s and A^u, since in the inference stage, it is expected to have the $A^u \rightarrow Y^u$ relation. However, existing datasets only have class labels that refer to actions in the video and unfortunately do not have the additional context to describe the video from different perspectives. For example, the CUB [39] dataset, in the domain of still images, offers 312 binary attributes and 15 part locations, as well as class labels, while the previously existing datasets for the video domain only have class labels. Due to the necessity of such a dataset for the low-data-based video classification problem, we present the HVU-LD dataset for the gap in this research domain.

3.3. HVU-LD Building Steps

This subsection outlines the methodology used to construct the HVU-LD dataset, focusing on two key aspects. Firstly, the importance of including a variety of classes in each meta-set is discussed. Secondly, 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 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 [26], 100 classes for meta-training, meta-validation, and meta-testing were selected from the Kinetics [9] 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 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 extracts semantic features of all possible classes over their videos' detailed attributes using Fasttext [37]. Then, according to the semantic features, the K-Means clustering algorithm is applied. Since the aim is building a semantically distant group of three meta-sets, we set the 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.

Algorithm 1 Meta-set Class Selection Algorithm

CLASS_SELECTOR(*classNames*)

classFeatures ← *empty list*

Extract semantic feature from each className using Fasttext method

for all *className* ∈ *classNames* **do**
classFeature ← *Fasttext*(*className*)
classFeatures.add(*classFeature*)
end for

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

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

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

According to the clustering-based algorithm, we analyzed the semantic distances between classes in the meta-sets of both our proposed dataset and existing 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. We repeated this process 10 times and present the scores in Table 2. In addition, the intra-class and inter-class distances of other existing datasets [7,8] 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.

Table 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 [45]	1.284	0.2977	4.313
HMDB-51 [45]	1.796	0.564	3.184
Kinetics-100 [26]	1.144	1.145	0.999
Sm2Sm-100 [26]	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**

4. SS-CADA-VAE Model

Low-data metric learning has received significant attention in various computer vision research areas, such as image classification [18,19] and object detection [46,47]. In this study, we propose a model named SS-CADA-VAE, which includes two single-layer variational autoencoder networks for visual and textual modalities. The model estimates a distribution for each class using these networks, based on the support samples, and conducts classification by comparing the query set samples with the estimated class distributions. Our model employs class distribution to build class representation, which differs from point estimation-based strategies [18,19,48], to mitigate overfitting concerns [49].

4.1. Variational Inference and 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 [50]. In particular, the posterior distribution of a latent variable model is obtained by the Bayes rule, which is given in Equation (1):

$$p_\phi(z|x) = \frac{p_\phi(x|z) \cdot p_\phi(z)}{p_\phi(x)} \ , \tag{1}$$

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 (2):

$$p_\phi(z|x) \approx q_\theta(z|x) \ , \tag{2}$$

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 (3):

$$ELBO = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p_\theta(z)) \ , \tag{3}$$

The objective function in Equation (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. Our Method

Our method, SS-CADA-VAE, uses CADA-VAE [11] as a baseline study. While our method shares the same network architecture and cross- and distribution alignment procedures with the baseline study, it proposes a semi-supervision module that has a non-parametric classifier and a class distribution enhancement mechanism that helps to improve the non-parametric classifier's performance. More specifically, that non-parametric metric-learning-based classifier estimates and then aggregates the posterior distribution of support set samples for each class and calculates the similarity between the query set samples' posterior distributions with the aggregated distributions, which are calculated from the support set samples, using KL-divergence. Besides that, the distribution enhancement mechanism continuously enhances the estimated posterior distribution during the training stage to obtain more separable aggregated posterior distributions, thereby reducing the inter-class distance.

4.3. 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 [11], 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:

$$\begin{aligned} \mathcal{L}_{CA_1} &= \sum_{i}^{N} \mid x_{visual}^{(i)} - Dec_{visual}(h_{textual}) \mid \\ \mathcal{L}_{CA_2} &= \sum_{i}^{N} \mid x_{textual}^{(i)} - Dec_{textual}(h_{visual}) \mid \\ \mathcal{L}_{CA} &= \mathcal{L}_{CA_1} + \mathcal{L}_{CA_2} \ , \end{aligned} \tag{4}$$

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 [51]. The Wasserstein distance-based distribution alignment loss can be given as:

$$\mathcal{L}_{DA} = \sum_{i}^{N} (|| \mu_{visual} - \mu_{textual} ||_2^2 \ + \ || \Sigma_{visual}^{1/2} - \Sigma_{textual}^{1/2} ||_{Frobenius}^2)^{1/2} \ , \tag{5}$$

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 (5) uses μ and Σ parameters, which are estimated from Enc_{visual} and $Enc_{textual}$.

4.4. NMC: Non-Parametric Metric-Learning Classifier Module

Metric-learning-based low-data learning algorithms [18,19,48] 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 [18,19,48], 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.

4.4.1. 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 $[\mu_{visual}, \sigma^2_{visual}]$ 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:

$$\mu = (\frac{1}{N})\sum_{i}^{N}(\mu^{(i)}),\ \sigma^2 = \sum_{i}^{N}((\frac{1}{N^2})\cdot\sigma^{(i)^2})\ , \tag{6}$$

According to Equation (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 $[\mu, \sigma^2]$ 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.

4.4.2. DDC: Distribution Distance-Based Classification

After estimating the class distributions as described in Section 4.4.1, 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:

$$\mathcal{D}_{\mathcal{KL}}(p||q) = \int_x p(x)\cdot\log(\frac{p(x)}{q(x)})\ . \tag{7}$$

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

$$\mathcal{D}_{\mathcal{KL}}(p||q) = \frac{1}{2}[\log(\frac{|\Sigma_q|}{|\Sigma_q|}) - 1 + \quad (\mu_p - \mu_q)^T\,\Sigma_q^-1\,(\mu_p - \mu_q) + tr\Sigma_q^-1\Sigma_p]\ , \tag{8}$$

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 (8) as follows:

$$Pr_\theta(y = m \mid \mathcal{N}_i(\mu_i, \sigma_i^2)) = \frac{\exp(\mathcal{D}_{\mathcal{KL}}(\mathcal{N}_i||\mathcal{N}_m))}{\exp(\sum_{j=1}^{n} \mathcal{D}_{\mathcal{KL}}(\mathcal{N}_i||\mathcal{N}_j))} \ , \tag{9}$$

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

$$\hat{y} = \arg\max_{m \in n} Pr_\theta(y = m|\mathcal{N}_i(\mu_i, \sigma_i^2)) \ , \tag{10}$$

4.5. SDEM: Supervision-Based Distribution Enhancement Module

Variational autoencoder (VAE) models are included in the class of unsupervised generative models [50]. As the size of existing datasets has grown enormously, labeling and performing analysis have become major hurdles. In this context, a semi-supervised learning strategy helps to understand the entire dataset by using a limited number of labeled data samples. More precisely, by using the limited labeled data, its leveraging power is used in many problems [52,53]. At the same time, a semi-supervised learning strategy under variational Bayesian methods has been investigated and has promising results [54].

Estimating the class distributions from support set samples from the low numbers of data or query set sample-based distribution estimation has some drawbacks and needs to be refined. The most obvious drawback is the lack of representation power and bias problem [55].

To alleviate this problem, we use the classification result, which is described in Section 4.4.2, to update the baseline model. As presented in Figure 2, we modify the base variational autoencoder architecture to the semi-supervised variational autoencoder type. In this context, after yielding a latent feature vector, z, in the training section, our model minimizes the negative log-probability $\mathcal{L}_{SS} = -\log(Pr_\theta(y = m \mid \mathcal{N}_i(\mu_i, \sigma_i^2)))$ for the ground truth class m using the Adam optimizer. By using that supervision objective, estimated class distributions are regularized and forced to more precise estimations.

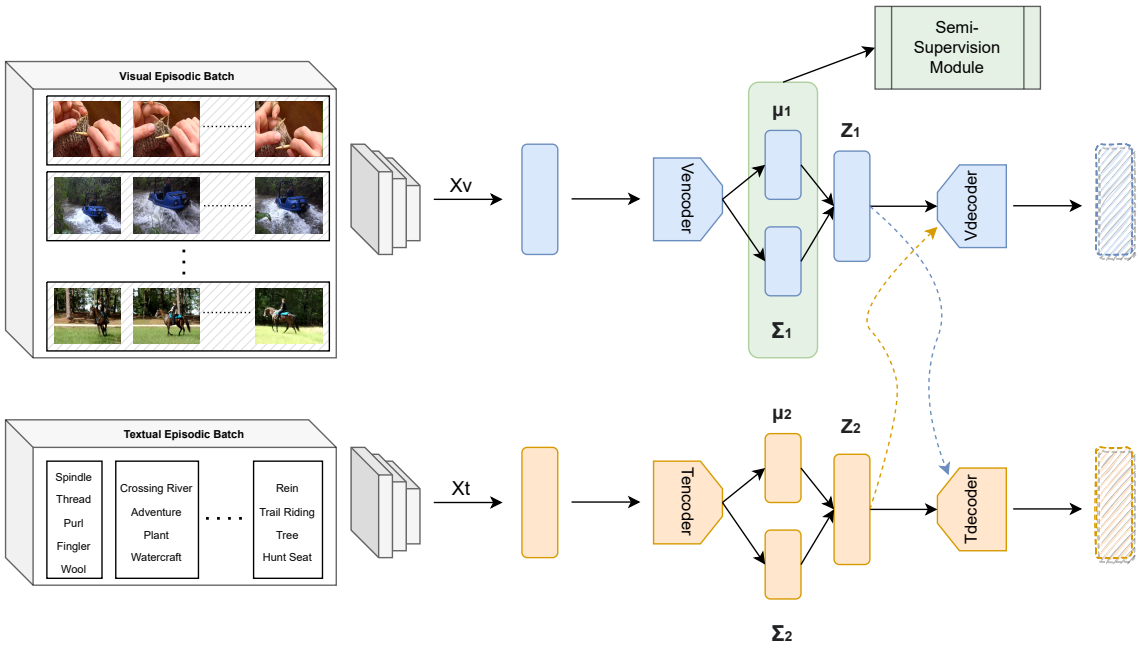


Figure 2. SS-CADA-VAE model with semi-supervision module (SSM). SSM includes two regularization terms to enhance visual distribution parameters in terms of decreasing intra-class distance and increasing inter-class distance. In addition to regularization, SSM includes distribution similarity-based classification operation, which is used in both the few-shot setting and zero-shot setting.

4.6. SS-CADA-VAE Loss

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

$$\mathcal{L}_{SS-CADA-VAE} = \mathcal{L}_{VAE} + \alpha\mathcal{L}_{CA} + \beta\mathcal{L}_{DA} + \gamma\mathcal{L}_{SS} \ , \quad (11)$$

In Equation (11), α, β, and γ refer the weights of the corresponding objective functions.

4.7. Implementation Details

CADA-VAE consists of two encoders and two decoders for each modality and each has one hidden layer. Owing to what the CADA-VAE study proposed for the still-image domain, the input for that model required some adaptations for the video-domain adaptation. In this context, since the video domain has more than one image, unlike the still-image domain in a single sample, input images, also known as frames, are extracted from the video, and then using different feature extractors, visual feature vectors are computed. As a last step, all these visual feature vectors are averaged to create a video-level feature vector. Thus, without a substantial change in the CADA-VAE architecture, it is made applicable to video-domain usage. The overall architecture of our model is illustrated in Figure 2.

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 [56] and visual transformer [57]. 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 for each video in uniform order and then resized to 224×224. 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 [58] and Bert [59]. 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.

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}. And 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.

Finally, all evaluations, model weights, complete code, and detailed information of the HVU-LD dataset are publicly available. Also, the model and all other components were implemented in the PyTorch framework.

5. Experiments

We 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 (3 classes), 5-way (5 classes), and 10-way (10 classes) scenarios. In the evaluation stage, we repeat each test result 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 [11] 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 4.4.2, to compare the improvements made by the SS-CADA-VAE more sharply. Behind all those combinations, we also compare the visual and textual feature extraction methods to investigate their impacts on low-data-based video classification.

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. Impact of Visual Feature Methodologies and Distribution Enhancement Module for Few-Shot Video Classification

In this analysis section, we present the impact of visual-level feature extraction methods and distribution enhancement module, $\mathcal{L}_{SS}$, on the low-data-based video classification problem. To produce a fair comparison, textual-level features are set as fixed with ELMO features. In this manner, experiments are divided into two groups in terms of the selected hyper-parameters. The first model type, *BL*, refers to the baseline model where γ is declined, and the second model type, *I-BL*, is the improved baseline model that includes the $\mathcal{L}_{SS}$ where $\gamma = 1$. The results for *5-way* few-shot learning are presented in Table 3.

Table 3. 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

The analysis of few-shot learning classification can be carried out in two ways: first, the performance of visual-level feature extraction methods without using any additional regularization, which is called the baseline study; second, the contribution of the regularization method $\mathcal{L}_{SS}$. Also, the number of samples in the support-set, k, is analyzed when making the performance comparison.

For the first question, for each hyper-parameter group (BL, I-BL), the ViT-based visual-level feature-oriented models outperform the Densenet201-based models' classification accuracy scores for each *k-shot* group in the few-shot learning problem. On average, the performance gap between visual feature extractors starts at 14.17% when only one sample is present in the support set ($k = 1$). In the end, the gap between the baseline models, BL CNN and BL ViT, narrows to 5.63%. These results suggest the importance of visual-level feature extraction methodologies in the few-shot learning problem.

When analyzing the impact of $\mathcal{L}_{SS}$ on the few-shot learning problem, BL and I-BL algorithms are compared while taking their visual-level feature extraction methodologies into consideration. On the Densenet-201 side, BL CNN and I-BL CNN, our newly proposed regularization mechanism, improve the classification performance overall by 4.21%. However, especially in the 1-shot setting, which is the crucial and main theme of the few-shot learning problem, improvement is 5.86%. It is important to note that this improvement has been achieved without using any cumbersome network changes or additional network parameter overloading. One noteworthy detail is that using $\mathcal{L}_{SS}$ regularization decreases the performance gap of Densenet-201 and ViT for all *k-shot* settings, while using nearly quadruples the efficient neural network architecture. On the ViT side, the few-shot classification accuracy of the ViT models improved by an average of 2.81% across all *k-shot* settings, and like in the Densenet-201 side analysis, the improvement is nearly 6% in the *5-way 1-shot* setting, which is a particularly important and challenging setting in real-world applications.

Another important analysis is based on the number of classes, *n-way*, in the support set. A general intuition is that when the number of classes increases, the few-shot classification problem becomes more complex and challenging. To evaluate the effect of this, Table 4 includes the results when k is fixed to 1, and n goes from 3 to 10, where k is the number of samples in the support set and n is the number of classes.

As shown in Table 4, $\mathcal{L}_{SS}$ regularization improves both the Densnet-201-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.

Table 4. 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

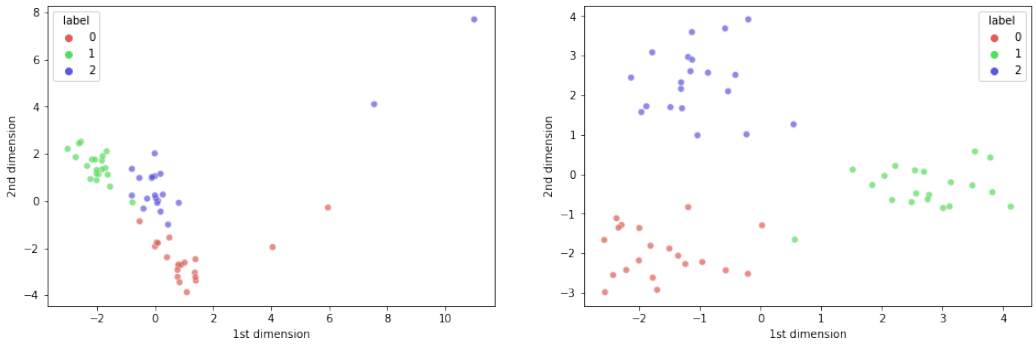
Also, we have analyzed the impact of $\mathcal{L}_{SS}$ for different levels. In this context, estimated distributions for classes have a crucial value in performing distribution distance-based classification (DDC), as shown in Section 4.4.2. In this manner, we analyze the $\mathcal{L}_{SS}$ objective to understand its impacts on low-data-based video classification. In this context, Table 5 shows the effect of SDEM, as shown in Section 4.5, on the estimated distribution parameters. In this experiment, we set the other hyper-parameters as fixed to observe the impact clearly.

Table 5. 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).

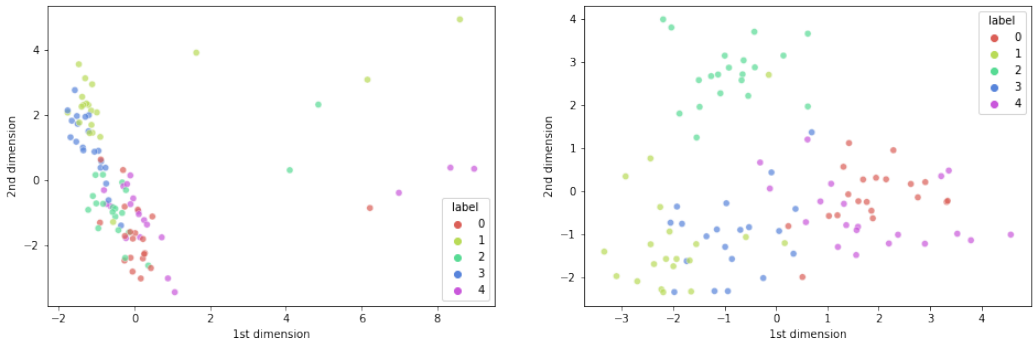
	1 Shot	3 Shot	5 Shot	10 Shot	Average
$\gamma = 0.0$ (BL)	51.42	71.38	77.25	82.88	70.73
$\gamma = 0.25$	60.91	78.54	82.68	85.77	76.97
$\gamma = 0.50$	63.70	77.96	82.10	84.80	77.14
$\gamma = 0.75$	47.44	66.39	72.20	78.03	66.01
$\gamma = 1.0$	57.28	76.23	81.22	85.03	74.94

As observed in Table 5, the estimated distribution parameters become more separable and distinguishable when the $\mathcal{L}_{SS}$ objective function is used. For example, in the *5-way 1-shot* setting, whereas the baseline model that does not have the $\mathcal{L}_{SS}$ objective function achieves 51.42% accuracy, our model, SS-CADA-VAE, has 63.70% accuracy when the coefficient of $\mathcal{L}_{SS}$ equals 0.5. Supplementary to the *5-way* setting, Figure 3 shows the impact of SDEM, as shown in Section 4.5, for the *3-way* and *5-way* setting visually. In the visualization stage, videos are encoded with Enc_{visual}, then their latent vectors' dimensions are reduced to two dimensions for visualization.

As a result of this ablation study, according to both Table 5 and Figure 3, using the $\mathcal{L}_{SS}$ objective in the training stage improves the classification performance dramatically by increasing the inter-class distance and preserving the intra-class distance.



(**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 3. The top-left and bottom-left figures 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 enhancement module *SDEM*. The top-right and bottom-right figures are 2D projections of the samples that are used in the left figures using the distribution enhancement module *SDEM*.

5.2. Impact of Visual and Textual Feature Methodologies and Distribution Enhancement Module for Zero-Shot Video Classification

In this analysis section, we present our experiments of zero-shot video classification. As the name implies, in the zero-shot learning setting, a class in the support set has no example, $k = 0$. Under this restriction, the video is classified by the collaboration of textual-level features and visual-level features.

In this context, we first experimented with the baseline model which does not have the regularization method, γ, then the impact of regularization methods was investigated. In this section, we analyzed the different textual-level feature extraction methodologies, ELMO and Bert, with the visual-level features, which were obtained by Densenet-201.

Table 6 summarizes the results of our zero-shot learning classification experiments using different types of textual-level features. As expected, the baseline performances for both textual-level features show a similar trend to their contextualized power in other

downstream tasks [60]. However, when the $\mathcal{LSS}$ regularization objective is applied, we observe a significant improvement in zero-shot classification performance. Specifically, the use of $\mathcal{LSS}$ 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.

Table 6. 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	52.19

5.3. Few-Shot Classification with HMDB-51

Our newly proposed low-data-based multi-modal video classification dataset is the first dataset that contains pairs of videos and attributes that capture multiple facets of the video content. Thus, future studies in this field will benefit from this dataset and be able to provide more real-world-compatible algorithms.

Therefore, we proposed a variational-autoencoder-based baseline study to evaluate this dataset and serve as a starting point for future studies. However, the generalization ability of the basic 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 [11], 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 [45], 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 7.

The results of the HMDB-51 dataset, which are presented in Table 7, 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 7. 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 [28]	54.5	65.7	68.0	-
HYRSM [45]	60.3	71.7	76.0	-
STRM [61]	-	-	81.30	-
FSVC-ATA [62]	65.78	-	82.27	-
BL [1]	42.92	64.47	71.55	79.02
I-BL (ours) [2]	60.01	**75.47**	**84.87**	**89.14**

[1] Refers to our baseline study. [2] Refers to our improved study.

6. Conclusions

In this paper, we address two different crucial points in the low-data-based video classification problem. First, we have proposed a new dataset, HVU-LD, that has detailed textual attributes. Another important role of this dataset will be to ensure consistency in the benchmarking of low-data-based video classification studies in the future, because, while the image domain has such a dataset, the video domain lacks this kind of dataset. Second, we have experimented with a baseline study on this dataset and proposed a non-parametric classifier and the class distribution enhancement mechanism to improve the baseline study's performance. And our new proposed class distribution enhancement mechanism's results surpass the baseline study's performance both for few-shot learning and zero-shot learning video classification settings.

Besides the tests on our newly proposed dataset, we have conducted the uni-modal few-shot learning-based video classification test on the previously existing benchmark dataset and compared our methods, which are the baseline and the improved baseline methods, between themselves and with other state-of-the-art studies. As a result of these experiments, the boost effect of our regularization technique on the baseline method has been revealed for both our proposed dataset and the HMDB-51 dataset.

In the future, we plan to extend our study to cover the generalized zero-shot video classification problem, as HVU-LD is suitable for addressing this research need. Also, end-to-end video-dedicated vision transformer-based studies will be conducted and their usages will be analyzed.

Author Contributions: Conceptualization, E.C. and M.E.K.; methodology, E.C. and M.E.K.; software, E.C.; validation, E.C. and M.E.K.; formal analysis, E.C. and M.E.K.; investigation, E.C. and M.E.K.; resources, E.C. and M.E.K.; data curation, E.C.; writing—original draft preparation, E.C. and M.E.K.; writing—review and editing, E.C. and M.E.K.; visualization, E.C. and M.E.K.; supervision, M.E.K.; project administration, M.E.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

In this section, an example of a meta-learning test episode is given in Figure A1. According to the given episode, it shows that each episode has three different pieces of information. One is the visual support set, another is the query set, which consists of videos' visual components, and the last one is the textual support set. There are different classes: Battle rope training, Bodybuilding, and Push up. In the visual support set, each class has three videos for the learning stage, and five different videos are presented for the testing stage in the query set. Besides that, in the zero-shot learning problem, as the problem implies, instead of the visual support set, the textual support set is used for the learning stage; some of the textual descriptions are omitted for the sake of simplicity, but thanks to [10], each video has detailed textual descriptions.

While the visual support set is used to construct class distributions in the few-shot learning stage, in the zero-shot learning stage, textual support set is used to construct class distributions. Calculated class distributions are compared with the query set samples' distributions for any learning stages to perform low-data-based video classification.

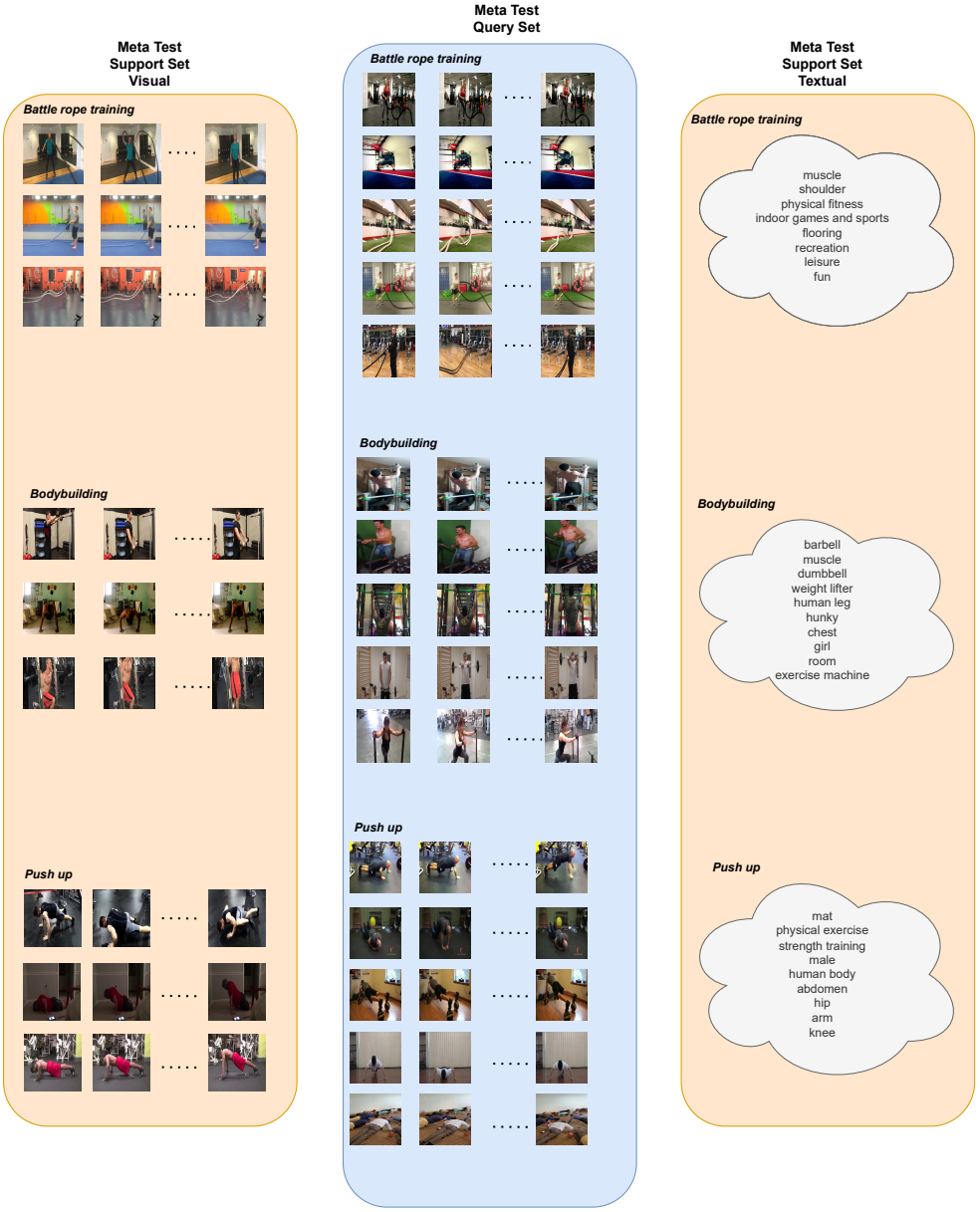


Figure A1. An episode of a meta-learning testing stage with support and query set components, in addition to the visual and textual information.

References

1. Girdhar, R.; Carreira, J.; Doersch, C.; Zisserman, A. Video action transformer network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 244–253.
2. Howard, J.P.; Tan, J.; Shun-Shin, M.J.; Mahdi, D.; Nowbar, A.N.; Arnold, A.D.; Ahmad, Y.; McCartney, P.; Zolgharni, M.; Linton, N.W.; et al. Improving ultrasound video classification: An evaluation of novel deep learning methods in echocardiography. *J. Med. Artif. Intell.* **2020**, *3*, 4. [CrossRef] [PubMed]
3. Xu, H.; Gao, Y.; Yu, F.; Darrell, T. End-to-end learning of driving models from large-scale video datasets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2174–2182.
4. Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, P.; Toderici, G.; Varadarajan, B.; Vijayanarasimhan, S. Youtube-8m: A large-scale video classification benchmark. *arXiv* **2016**, arXiv:1609.08675.
5. Lee, J.; Abu-El-Haija, S. Large-scale content-only video recommendation. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 987–995.
6. Gal, Y.; Islam, R.; Ghahramani, Z. Deep bayesian active learning with image data. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1183–1192.

7. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2556–2563.
8. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.
9. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The kinetics human action video dataset. *arXiv* **2017**, arXiv:1705.06950.
10. Diba, A.; Fayyaz, M.; Sharma, V.; Paluri, M.; Gall, J.; Stiefelhagen, R.; Gool, L.V. Large scale holistic video understanding. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 593–610.
11. Schonfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; Akata, Z. Generalized zero-and few-shot learning via aligned variational autoencoders. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8247–8255.
12. Wang, Y.; Yao, Q.; Kwok, J.T.; Ni, L.M. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.* **2020**, *53*, 1–34. [CrossRef]
13. Lake, B.M.; Salakhutdinov, R.; Tenenbaum, J.B. Human-level concept learning through probabilistic program induction. *Science* **2015**, *350*, 1332–1338. [CrossRef] [PubMed]
14. Zhang, H.; Zhang, J.; Koniusz, P. Few-shot learning via saliency-guided hallucination of samples. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2770–2779.
15. Li, A.; Luo, T.; Xiang, T.; Huang, W.; Wang, L. Few-shot learning with global class representations. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9715–9724.
16. Nakamura, A.; Harada, T. Revisiting fine-tuning for few-shot learning. *arXiv* **2019**, arXiv:1910.00216.
17. Hu, S.X.; Li, D.; Stühmer, J.; Kim, M.; Hospedales, T.M. Pushing the limits of simple pipelines for few-Shot learning: External data and fine-tuning make a difference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9068–9077.
18. Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. Matching networks for one shot learning. In Proceedings of the Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016.
19. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4080–4090.
20. Pahde, F.; Puscas, M.; Klein, T.; Nabi, M. Multimodal prototypical networks for few-shot learning. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Virtual, 5-9 January 2021; pp. 2644–2653.
21. Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; Akata, Z. Attribute prototype network for zero-shot learning. In Proceedings of the Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; pp. 21969–21980.
22. Wu, J.; Tian, X.; Zhong, G. Supervised Contrastive Representation Embedding Based on Transformer for Few-Shot Classification. *J. Phys. Conf. Ser.* **2022**, *2278*, 12–22. [CrossRef]
23. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10012–10022.
24. Hubert Tsai, Y.H.; Huang, L.K.; Salakhutdinov, R. Learning robust visual-semantic embeddings. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3571–3580.
25. Chen, S.; Hong, Z.; Liu, Y.; Xie, G.S.; Sun, B.; Li, H.; Peng, Q.; Lu, K.; You, X. Transzero: Attribute-guided transformer for zero-shot learning. *Proc. Aaai Conf. Artif. Intell.* 2022, *36*, 330–338. [CrossRef]
26. Zhu, L.; Yang, Y. Compound memory networks for few-shot video classification. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 751–766.
27. Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. The "something something" video database for learning and evaluating visual common sense. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5842–5850.
28. Cao, K.; Ji, J.; Cao, Z.; Chang, C.Y.; Niebles, J.C. Few-shot video classification via temporal alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10618–10627.
29. Muller, M. *Information Retrieval for Music and Motion*; Springer: Berlin/Heidelberg, Germany, 2007.
30. Jin, R.; Wang, X.; Wang, G.; Lu, Y.; Hu, H.M.; Wang, H. Embedding adaptation network with transformer for few-shot action recognition. In Proceedings of the Asian Conference on Machine Learning, İstanbul, Turkey, 11–14 November 2023; pp. 515–530.
31. Wang, X.; Zhang, S.; Cen, J.; Gao, C.; Zhang, Y.; Zhao, D.; Sang, N. CLIP-guided prototype modulating for few-shot action recognition. *J. Comp. Vis.* **2023**, 1–14. [CrossRef]
32. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763.

33. Wang, X.; Zhang, S.; Yuan, H.; Zhang, Y.; Gao, C.; Zhao, D.; Sang, N. Few-shot Action Recognition with Captioning Foundation Models. *arXiv* **2023**, arXiv:2310.10125.
34. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
35. Fabian, C.H.; Victor, E.; Bernard, G.; Juan, C.N. ActivityNet: A large-scale video benchmark for human activity understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 961–970.
36. Brattoli, B.; Tighe, J.; Zhdanov, F.; Perona, P.; Chalupka, K. Rethinking zero-shot video classification: End-to-end training for realistic applications. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4613–4623.
37. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
38. Lin, C.C.; Lin, K.; Wang, L.; Liu, Z.; Li, L. Cross-modal representation learning for zero-shot action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 19978–19988.
39. Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; Perona, P. *Caltech-UCSD Birds 200*; Technical Report; California Institute of Technology: Pasadena, CA, USA, 2010.
40. Patterson, G.; Hays, J. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2751–2758.
41. Lampert, C.H.; Nickisch, H.; Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In Proceedings of the IEEE conference on computer vision and pattern recognition, Miami, FL, USA, 20–25 June 2009 ; pp. 951–958.
42. Xian, Y.; Lampert, C.H.; Schiele, B.; Akata, Z. Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2251–2265. [CrossRef] [PubMed]
43. Google Cloud Vision API. Available online: https://cloud.google.com/vision (accessed on 1 February 2024).
44. Sensifai Video Tagging API. Available online: https://sensifai.com (accessed on 1 February 2024).
45. Wang, X.; Zhang, S.; Qing, Z.; Tang, M.; Zuo, Z.; Gao, C.; Jin, R.; Sang, N. Hybrid relation guided set matching for few-shot action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 19948–19957.
46. Karlinsky, L.; Shtok, J.; Harary, S.; Schwartz, E.; Aides, A.; Feris, R.; Giryes, R.; Bronstein, A.M. Repmet: Representative-based metric learning for classification and few-shot object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5197–5206.
47. Shen, J.; Wang, H.; Zhang, A.; Qiu, Q.; Zhen, X.; Cao, X. Model-agnostic Metric for zero-shot learning. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 786–795.
48. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.; Hospedales, T.M. Learning to compare: Relation network for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1199–1208.
49. Zhang, J.; Zhao, C.; Ni, B.; Xu, M.; Yang, X. Variational few-shot learning. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1685–1694.
50. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
51. Givens, C.R.; Shortt, R.M. A class of wasserstein metrics for probability distributions. *Mich. Math. J.* **1984**, *31*, 231–240. [CrossRef]
52. Yalniz, I.Z.; Jégou, H.; Chen, K.; Paluri, M.; Mahajan, D. Billion-scale semi-supervised learning for image classification. *arXiv* **2019**, arXiv:1905.00546.
53. Li, A.; Yuan, P.; Li, Z. Semi-supervised object detection via multi-instance alignment with global class prototypes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9809–9818.
54. Kingma, D.P.; Mohamed, S.; Jimenez Rezende, D.; Welling, M. Semi-supervised learning with deep generative models. *Adv. Neural Inf. Process.* **2014**, *27*, 3581–3589.
55. Bibby, J. Some basic theory for statistical inference. *Math. Gaz.* **1980**. *64*, 138–138. [CrossRef]
56. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
57. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
58. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
59. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
60. Reimers, N.; Schiller, B.; Beck, T.; Daxenberger, J.; Stab, C.; Gurevych, I. Classification and clustering of arguments with contextualized word embeddings. *arXiv* **2019**, arXiv:1906.09821.

61. Thatipelli, A.; Narayan, S.; Khan, S.; Anwer, R.M.; Khan, F.S.; Ghanem, B. Spatio-temporal relation modeling for few-shot action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 19958–19967.
62. Nguyen, K.D.; Tran, Q.H.; Nguyen, K.; Hua, B.S.; Nguyen, R. Inductive and transductive few-shot video classification via appearance and temporal alignments. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 471–487.