

# Optimizing IoT Video Data: Dimensionality Reduction for Efficient Deep Learning on Edge Computing

David Ortiz-Perez<sup>1</sup>, Pablo Ruiz-Ponce<sup>1</sup>, David Mulero-Pérez<sup>1</sup>,  
Manuel Benavent-Lledo<sup>1</sup>, Javier Rodriguez-Juan<sup>1</sup>,  
Hugo Hernandez-Lopez<sup>1</sup>, Anatoli Iarovikov<sup>1</sup>, Srdjan Krco<sup>2</sup>,  
Daliborka Nedic<sup>2</sup>, Dejan Vukobratovic<sup>3</sup>, Jose Garcia-Rodriguez<sup>1,\*</sup>

<sup>1</sup> *Department of Computer Science and Technology,  
University of Alicante, Spain*

<sup>2</sup> *DunavNet, Novi Sad, Serbia*

<sup>3</sup> *Faculty of Technical Sciences, University of Novi Sad, Serbia*

\* Correspondence: [jgarcia@dtic.ua.es](mailto:jgarcia@dtic.ua.es)

## Abstract

The rapid loss of biodiversity significantly impacts birds' environments and behaviors, highlighting the importance of analyzing bird behavior for ecological insights. With the growing adoption of Machine Learning (ML) algorithms in the Internet of Things (IoT) domain, edge computing has become essential to ensure data privacy and enable real-time predictions by processing high-dimensional data, such as video streams, efficiently. This paper introduces a set of dimensionality reduction techniques tailored for video sequences based on cutting-edge methods for this data representation. These methods drastically compress video data, reducing bandwidth and storage requirements while enabling the creation of compact ML models with faster inference speeds. Comprehensive experiments on bird behavior classification in rural environments demonstrate the effectiveness of the proposed techniques. The experiments incorporate state-of-the-art deep learning techniques, including pre-trained video vision models, Autoencoders, and single-frame feature extraction. These methods demonstrated superior performance to the baseline, achieving up to a 6000-fold reduction in data size while reaching a classification accuracy of 60.7% on the Visual WetlandBirds Dataset and obtaining state-of-the-art performance on this dataset. These findings underline the potential of using dimensionality reduction to enhance the scalability and efficiency of bird behavior analysis.

**Keywords:** deep learning; spatio-temporal dimensionality reduction; bird behavior classification; edge computing; video processing

## 1 Introduction

The accelerating pace of global biodiversity loss necessitates efficient environmental management strategies to mitigate its impacts [1]. Bird behavior, closely influenced by envi-

ronmental conditions, weather, and surrounding ecosystems, serves as a valuable indicator of ecological health [2]. Analyzing bird behavior provides critical insights for researchers and ecologists, enabling the identification of behavioral changes and anomalies that may signal environmental shifts [3, 4]. As highly sensitive species, birds can detect subtle changes in their habitats, making them indispensable in monitoring and conservation efforts [5].

Video classification can be used for automating the study of bird behavior. However, this task can be resource-intensive, requiring the analysis of both visual and temporal data, which presents significant challenges in terms of time and effort. Additionally, it presents several challenges inherent to the task itself. The main one is the high dimensionality of the data the models have to process. This leads to a sparse data distribution due to the curse of dimensionality [6], requiring lots of samples to train a classifier without significant overfitting. Moreover, the high dimensionality significantly increases the number of trainable parameters in the classifiers, which in turn demands greater computational resources for both training and inference. In the context of bird behavior recognition, resources are often highly constrained due to deployment in rural environments.

To address these challenges, we present a study focused on reducing the dimensionality of video recordings by transforming them into smaller and more compact representations while preserving the most relevant features for classification. To achieve this, we employ various state-of-the-art deep learning techniques, including the employment of embeddings obtained from large video vision models, such as 3D Convolutional Neural Networks (3DCNNs) [7, 8] or Video Transformers [9, 10]. In addition, other techniques are proposed, including Autoencoders and extracting visual features from a single frame within the sequence. We present a comprehensive ablation study to identify the most effective methods and parameters. Our results demonstrate that the proposed methodologies can reduce video representations by a factor of 6000, while these compact representations outperform the baseline model. All experiments were carried out on the Visual WetlandBirds Dataset [11], yielding up to a 60.7% accuracy increase over the test set and presenting new state-of-the-art performance. Furthermore, they significantly reduce training time due to the smaller data size. In summary, our key contributions are as follows:

- We proposed three different approaches to dimensionality reduction for the specific task of action classification from videos of birds in natural environments. These methods can reduce the size of a video by up to 6000 times for much faster training and inference.
- Additionally, the new classifiers trained with the reduced representations outperformed the previously proposed baselines, achieving state-of-the-art results on this specific dataset. In this particular task, characterized by a limited number of samples, our reduced representation has proven more effective in capturing differences between classes.
- Finally, we performed an exhaustive ablation study of the different proposed methods to determine parameter selection and the importance of each of them. This process allowed us to establish the capabilities and limitations of each method.

The remainder of this paper is structured as follows: section 2 reviews the most relevant related work within this field. section 3 outlines the methodology employed in this study. section 4 presents the experimental results. section 5 discusses and interprets

the results obtained from this study. Finally, section 6 discusses the conclusions drawn from this work, along with potential directions for future research.

## 2 Related Work

Deep learning techniques excel at identifying and analyzing patterns in heterogeneous, high-dimensionality data. However, certain data modalities and scenarios benefit from reduced representations. For instance, video data present challenges due to their high spatial and temporal dimensionality. Similarly, in scenarios like the Internet of Things (IoT), where computational resources are limited and low inference times are critical, using reduced data representations leads to smaller models with lower computational costs and faster inference times.

In this context, dimensionality reduction involves mapping a set of input features to a smaller set while retaining meaningful information that can still be used for the same tasks as the original representation [12, 13, 14, 15]. Two main approaches are commonly discussed in the literature [12]: feature selection and feature extraction.

Feature selection focuses on identifying and retaining a subset of the original features that preserve the most information. These features remain in their original form, and the methods prioritize selecting them based on specific criteria. This approach is particularly useful in scenarios with low-sample, high-dimensionality tabular data, where noise and feature redundancy are prevalent.

In contrast, feature extraction transforms the initial features into a lower-dimensional representation, making it more suitable for unstructured data such as images and videos. Traditional linear approaches aim to achieve this transformation through linear methods. Examples include techniques based on variances and contribution ratios, such as Principal Component Analysis (PCA) [16, 17], Linear Discriminant Analysis (LDA) [18], and Factor Analysis (FA). Other linear feature extraction methods include Independent Component Analysis (ICA), Multi-Dimensional Scaling (MDS) [19], and Singular Value Decomposition (SVD) [20]. However, these linear methods are limited in their ability to address the inherent non-linearity and complexity of video data.

Non-linear feature extraction algorithms provide flexibility to handle complex, non-linear relationships in data that linear methods fail to capture effectively. Examples include Kernel Principal Component Analysis (KPCA) [21], which uses kernel functions to project data into higher-dimensionality spaces where linear separation is feasible, and Locally Linear Embedding (LLE) [22], which maintains local neighborhood relationships while embedding data into a low-dimensionality space. Isometric Mapping (ISOMAP) [23] is another prominent approach, extending classical MDS by preserving geodesic distances between data points, effectively capturing the manifold structure of complex data. Non-negative Matrix Factorization (NMF) [24] offers a part-based representation of data by imposing non-negativity constraints, making it highly interpretable in fields such as text mining and facial recognition. These non-linear methods provide improved capability to represent the complexity and inherent non-linear nature of video and image data compared to traditional linear approaches.

Dimensionality reduction using deep learning can be categorized into supervised [25], unsupervised, and semi-supervised methods. Supervised methods, such as Convolutional Neural Networks (CNNs) [26] and Transformer [27] models, effectively reduce feature dimensions by capturing both local and global features present in data. These methods

work particularly well for structured tasks where labeled data are available, enabling the extraction of key information while minimizing irrelevant details. On the other hand, unsupervised methods, like Deep Autoencoders [28, 29, 30, 31] and Deep Belief Networks (DBNs) [32], focus on compressing high-dimensional data into lower-dimensional representations without relying on labels, making them suitable for discovering hidden structures and patterns in unstructured data.

Semi-supervised approaches leverage the strengths of both supervised and unsupervised learning by combining unsupervised pre-training with labeled data to improve generalization. For instance, pre-training with unlabeled data helps the model learn meaningful representations, while fine-tuning with labeled data refines its accuracy. This combination is particularly valuable when labeled data are scarce, enhancing the model’s ability to generalize effectively. Overall, deep learning-based dimensionality reduction methods can handle the complexity and non-linearity of high-dimensionality data [33], providing efficient and meaningful feature representations for tasks like image, video, and sensor data analysis.

Dimensionality reduction is crucial for handling video data due to their high computational complexity and memory requirements and the challenge of extracting effective spatio-temporal features. Unlike static data, videos involve a temporal dimension, which requires specialized approaches for the efficient management of both spatial and temporal aspects [12]. Various methods have been developed to reduce dimensionality, making video data more manageable while maintaining the quality of extracted features. Techniques like the Spatio-temporal Prompting Network (STPN) [34], Regularized Deep Neural Networks (rDNNs) [35], and lightweight optimization for frame interpolation [36] exemplify how dimensionality reduction can enhance performance in video classification, detection, and segmentation tasks. These methods focus on reducing redundancy, preserving essential information, and improving computational feasibility.

Dimensionality reduction also finds applications beyond video classification, such as human action recognition, medical video analysis, and human detection. For action recognition, local CNN features are aggregated to create global representations, addressing GPU memory limitations while preserving performance [37]. In medical video analysis, particularly gastrointestinal endoscopy, hybrid feature extraction techniques reduce computational costs while retaining critical diagnostic information [38, 39]. Additionally, incremental Principal Component Analysis (PCA) has been used to improve human detection by reducing the dimensionality of CoHOG features [40]. Overall, dimensionality reduction plays an essential role in making video data analysis practical and improving efficiency, memory management, and accuracy across various video-related applications.

In the study of animal behavior [41, 42, 43] and, more specifically, in bird-related research [44, 45, 46, 47], videos serve as a primary data structure for analysis and insight generation. However, to the best of our knowledge, no prior work has applied dimensionality reduction techniques to this specific domain, making our approach the first to explore and integrate these methods for analyzing video data in the context of animal behavior and avian studies.

When considering datasets for training deep learning algorithms, there is a notable scarcity of resources focused on bird behavior recognition. For instance, the VB100 dataset [48], which consists of video recordings of birds in their natural environments, is annotated only with bird species and lacks labels for the actions being performed. Consequently, this dataset is not suitable for behavior recognition tasks. Similarly, the AnimalKingdom dataset [49] contains video recordings of various animals and includes

annotations of their corresponding actions. However, as this dataset encompasses a wide range of animal species, including but not limited to birds, it is not suitable for focused bird behavior recognition. The Visual WetlandBirds Dataset [11] provides video recordings exclusively of birds, making it the most appropriate dataset for this study. In addition, other datasets, such as Birds525 (<https://huggingface.co/datasets/chriamue/bird-species-dataset>, accessed on 18 January 2024), CUB-200-2011 [50], and NABirds [51], offer images of different bird species, but these datasets lack temporal or behavioral annotations. As a result, they are unsuitable for tasks requiring such information.

### 3 Materials and Methods

In this section, we present the methodology proposed for this study. The main scope of this study and the proposed pipeline are formulated to capture the most relevant information within video frames while converting videos into lower-dimensionality representations.

The proposed techniques aim to generate lower-dimensionality representations of the video input. Each technique includes a final Multi-Layer Perceptron (MLP) designed to classify and evaluate the effectiveness of the representations. An ablation study was conducted to identify the most suitable technique and the optimal parameters for training and the MLP. As a result, the configuration of the final MLP may vary depending on the employed method. All code developed for this project is publicly available in our GitHub repository (<https://github.com/3dperceptionlab/DimensionalityReductionBirdBehaviours>, accessed on 1 December 2024).

#### 3.1 Dataset

The Visual WetlandBirds Dataset [11] consists of videos of birds from the Valencian region, each depicting different actions. The dataset consists of 2765 video frames and is distributed into 1834 samples for training, 440 for validation, and 491 for testing. This distribution was designed taking into consideration the bird species in the video, as well as the action performed in it. Furthermore, this division was not limited to bird species and actions but also considered the original video source, which was segmented into smaller video frames. This approach ensures that videos are separated based on factors such as the recording camera, the day of recording, and the time of day. These videos are annotated with the bird species and the actions performed. The videos encompass 16 frames, with three color channels per frame and a resolution of  $224 \times 224$  pixels. This dataset establishes a baseline for the classification of bird behaviors using 3DCNNs and Video Transformers. The models trained end-to-end include ResNet3D, S3D, Video Swin Transformer, and MViTv2.

#### 3.2 Features

With the advent of deep learning technologies, a wide range of pre-trained models are available for various modalities and purposes. For instance, in Natural Language Processing (NLP), models such as BERT [52], LLaMA [53], and Gemini [54] have been developed. These models are trained on extensive datasets and can capture and encode the most rel-

evant information from an input, leveraging the knowledge acquired during their training phase.

Similarly, for video analysis, several large pre-trained models can be used for feature extraction and classification tasks. These models have been trained on the Kinetics Dataset [55], enabling them to learn general patterns that can be effectively leveraged for transfer to other tasks, for example, bird behavior classification. These models rely on diverse backbone architectures, such as 3DCNNs [56] and Video Transformers [57]. Examples of 3DCNNs include ResNet3D [7] and Video S3D [8], while Video Transformers include models like Video Swin Transformer [9] and MViTv2 [10].

The *Features* dimensionality reduction technique involves utilizing these models to extract internal representations of videos, generating embeddings that encapsulate essential features. These embeddings serve as input into an MLP for the final classification task. This architecture processes a sequence of 16 frames, each with dimensions of  $16 \times 3 \times 224 \times 224$ , and transforms it into a single embedding with a size of 400. This transformation achieves a dimensionality reduction of over 6000 times compared to the original video sequence, substantially decreasing the input size while preserving critical information. An overview of the proposed architecture is shown in Figure 1.

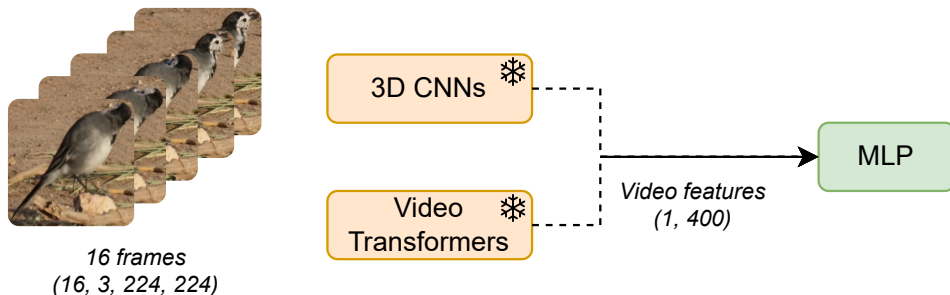


Figure 1: Overview of the Baseline Feature Method: The architecture utilizes the internal representations of the proposed models in conjunction with an MLP. Discontinued lines indicate that only one of the models is selected for use in each experiment. The snowflake symbol denotes a frozen model.

### 3.3 Autoencoder

An alternative methodology proposed in this study utilizes Autoencoders, designed to reduce the dimensionality of input data and reconstruct them from a latent space. This approach allows the model to compress the input into a lower-dimensionality representation while retaining its essential features. An Autoencoder is composed of two primary components: an encoder, which reduces the dimensionality of the input data, and a decoder, which reconstructs the input data to their original form [58, 31, 59].

During the training process, the Autoencoder learns to effectively represent the input data within the latent space and reconstruct them. Once trained, the decoder is removed from the pipeline. The latent space representation is used for downstream tasks. This study employs this latent space representation and an MLP to classify bird behaviors.

The spatio-temporal Autoencoder architecture is designed to process video data by utilizing 3D convolutions to capture spatial and temporal features. The encoder comprises four 3D convolutional layers that progressively reduce the spatial and temporal dimensions of the input. Each convolutional layer is followed by batch normalization

and ReLU activation to enhance convergence and improve model robustness. The output of the encoder is flattened and passed through a fully connected layer, compressing the data into a 1024-dimensional latent space. The decoder is structured to mirror the encoder, employing transposed 3D convolutions to reconstruct the video data from the latent representation. However, after the training process, the decoder is discarded, and representations from the latent space are used to reduce the dimensionality of video inputs. The learning rate for the training process is set to 0.0001, and the Adam optimizer is employed. Training is conducted over 100 epochs with a batch size of 32. The Mean Squared Error (MSE) loss function is utilized to minimize the pixel-wise difference between the input videos and their reconstructions in alignment with the Autoencoder’s objective. During training, the performance is periodically evaluated by saving the model state every 20 epochs, providing progress checkpoints, and facilitating recovery if needed.

This method allows for a significant reduction in data size, compressing a  $16 \times 3 \times 224 \times 224$  input into a single embedding of size 1024—a reduction of more than 2350 times the original data volume. Figure 2 provides an overview of the proposed architecture.

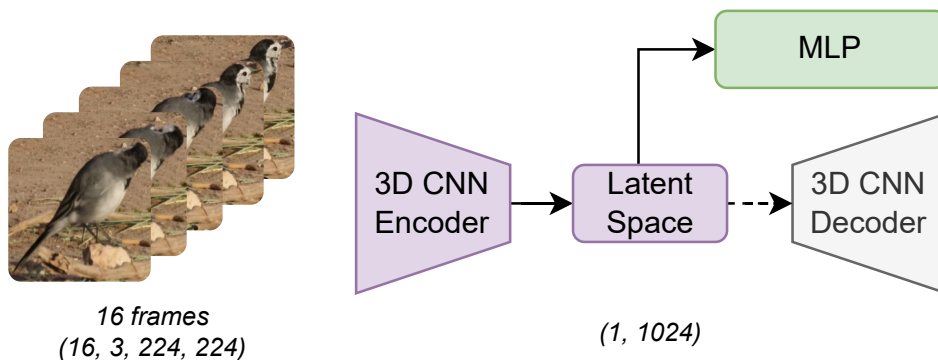


Figure 2: Overview of the Autoencoder Method: An Autoencoder is pre-trained to reconstruct video frames. Following the training process, the decoder is discarded, and the latent space representation is combined with an MLP. Discontinued lines indicate that the decoder is discarded after being trained, and only the latent space is used to classify the bird action.

### 3.4 Single Frame

The final dimensionality reduction technique implemented in this study involves using a single frame for the task, thereby discarding the temporal information contained within the video. The selected architecture processes a single frame, utilizing both the environmental context and the bird-specific features captured at that moment to perform the classification. We propose two primary approaches to extract a single frame from the video sequence. The first one is selecting the central frame of the sequence, while the second is computing the mean pixel values across the sequence to generate a representative frame. This selection serves as the first step in dimensionality reduction, reducing the input dimensions from  $16 \times 3 \times 224 \times 224$  to a single frame of  $3 \times 224 \times 224$ , effectively achieving a 16-fold reduction in size.

Following this initial dimensionality reduction step and drawing on methodologies used in previously proposed architectures, we employed pre-trained models such as 2D CNNs, Vision Transformers, the DINOv2 [60, 61] model, and a Histogram of Oriented Gradients (HOG) [62]. The 2D CNNs utilized include ResNet [63], MobileNet [64],

DenseNet [65], and VGG [66], while the implemented Vision Transformer models include ViT [67] and Swin Transformer [68]. The primary objective of this methodology is to leverage CNN kernels and transformer layers to extract meaningful patterns and features from images for classification tasks. The extent of dimensionality reduction achieved depends on the chosen model. For example, using ResNet, which delivers the most promising results among the 2D CNNs, we obtained embeddings of size 2048, achieving a reduction of over 1100 times. Other 2D CNNs produce varying reduction rates based on their internal representation sizes. For Vision Transformers and the DINOv2 model, the embedding size is 768, resulting in a reduction of more than 3100 times in size. HOG features were computed to extract complementary information [69, 70], with these features yielding a reduction size of 1,031,940, corresponding to a more modest reduction of slightly over two times. As in prior methodologies, the extracted representations were combined with an MLP for final classification. Figure 3 illustrates the overall architecture.

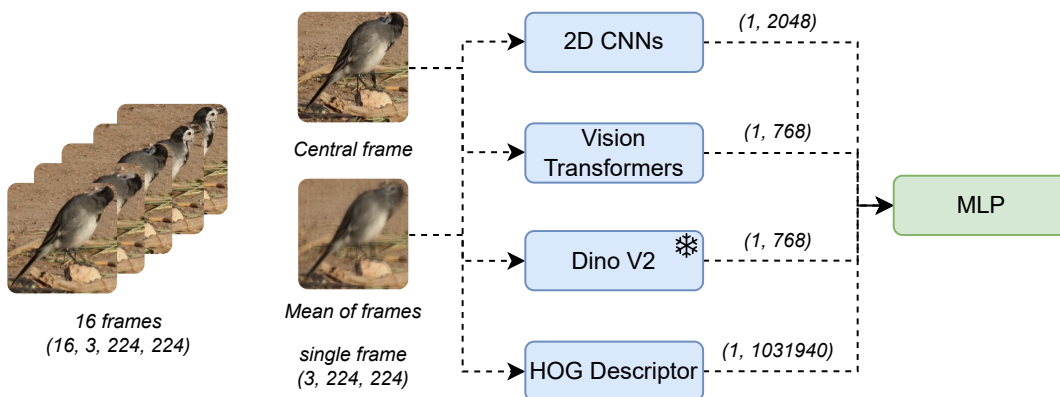


Figure 3: Overview of single-frame methods. The initial dimensionality reduction step involves selecting either the central frame or the mean pixel values of the frames. Subsequently, visual embeddings are obtained using 2D CNNs, Vision Transformers, or DinoV2. Finally, classification is performed using an MLP. Discontinued lines indicate that only one of the models and frames are selected for use in each experiment.

## 4 Results

In this section, we present quantitative experimentation with the methods previously introduced in this work. In table 1, we can observe the evaluation of the different proposed approaches and their reduction rates. By analyzing the results, we can observe that the method labeled *Features* is the best concerning reduction and metrics, even outperforming the original metrics obtained by the baseline.

Table 1: Quantitative evaluation of the proposed dimensionality reduction techniques, including their performance metrics and the reduction achieved relative to the original input size. Evaluations were conducted 5 times to mitigate the effects of randomness. The symbol  $\pm$  indicates the 95% confidence interval.

Model	Accuracy	Precision	Recall	F1	Reduction
Baseline	0.558	0.335	<b>0.438</b>	0.371	-
Features	<b>0.607</b> $\pm 0.004$	0.423 $\pm 0.02$	0.424 $\pm 0.009$	0.398 $\pm 0.05$	<b>6021</b>
Autoencoder	0.513 $\pm 0.002$	0.279 $\pm 0.25$	0.170 $\pm 0.001$	0.144 $\pm 0.003$	235
Single Frame (mean)	0.576 $\pm 0.004$	<b>0.442</b> $\pm 0.01$	0.419 $\pm 0.01$	<b>0.409</b> $\pm 0.004$	1176
Single Frame (central)	0.556 $\pm 0.007$	0.385 $\pm 0.006$	0.390 $\pm 0.01$	0.368 $\pm 0.004$	3136

The experiments demonstrate that the most relevant features were obtained by extracting embeddings from large pre-trained video models. These embeddings effectively capture high-level information from prior tasks, reducing the representation size by more than 6000 times while improving performance across various metrics. While this method is highly efficient for training, its benefits are more limited during the inference time due to the time required to extract the embeddings from the pre-trained models. The effectiveness of this method can be graphically observed in fig. 4.

The second-best method involves removing temporal information by processing only a single frame from each video sequence. We tested two strategies for selecting this single frame: The first was picking the central frame and calculating the mean frame pixel-wise across the entire sequence. The results indicate that using the mean frame leads to slightly better performance than using the central frame, likely because it retains some contextual information from the entire sequence. Features are extracted from the single frame using pre-trained image models, being a more lightweight process than the previous method in the video processing stage. As such, this could be the most suitable option in resource-constrained environments. Despite removing temporal information, this method achieves highly competitive performance metrics.

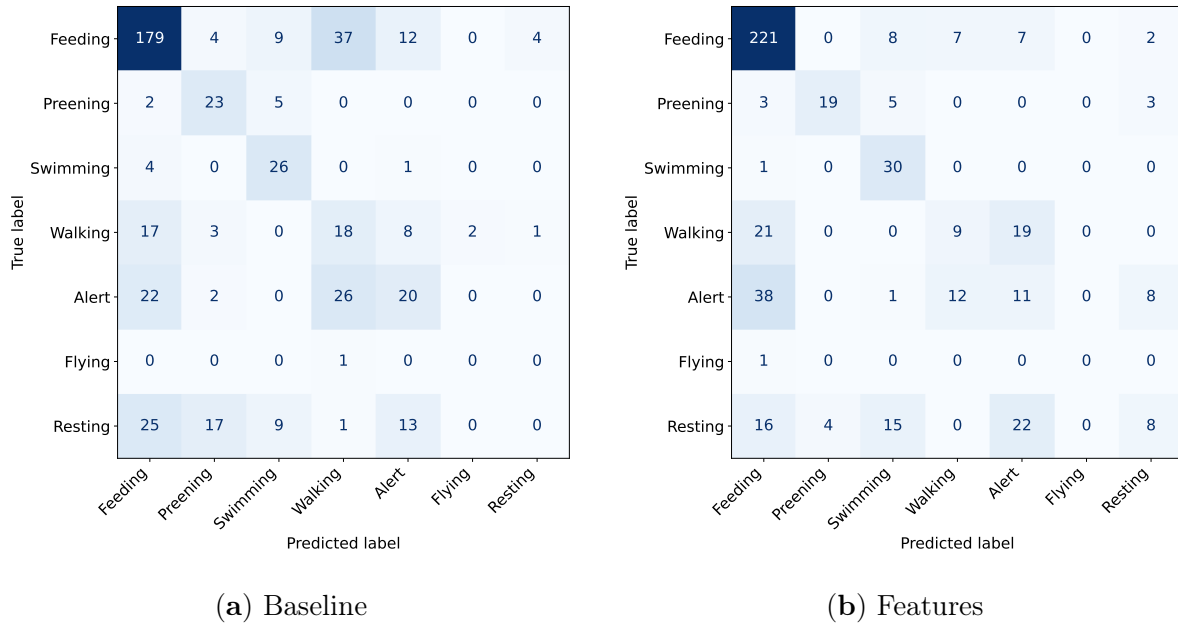


Figure 4: Confusion matrix comparing the baseline approach with the proposed best dimensionality reduction method on the Visual WetlandBirds Dataset test set.

Among the methods we proposed, the one based on Autoencoders performed the worst. Its lack of effectiveness can be attributed to overfitting, which occurred due to the limited availability of large datasets for this particular task. Figure 5 shows the learning curves for the loss functions for the different proposed methods.

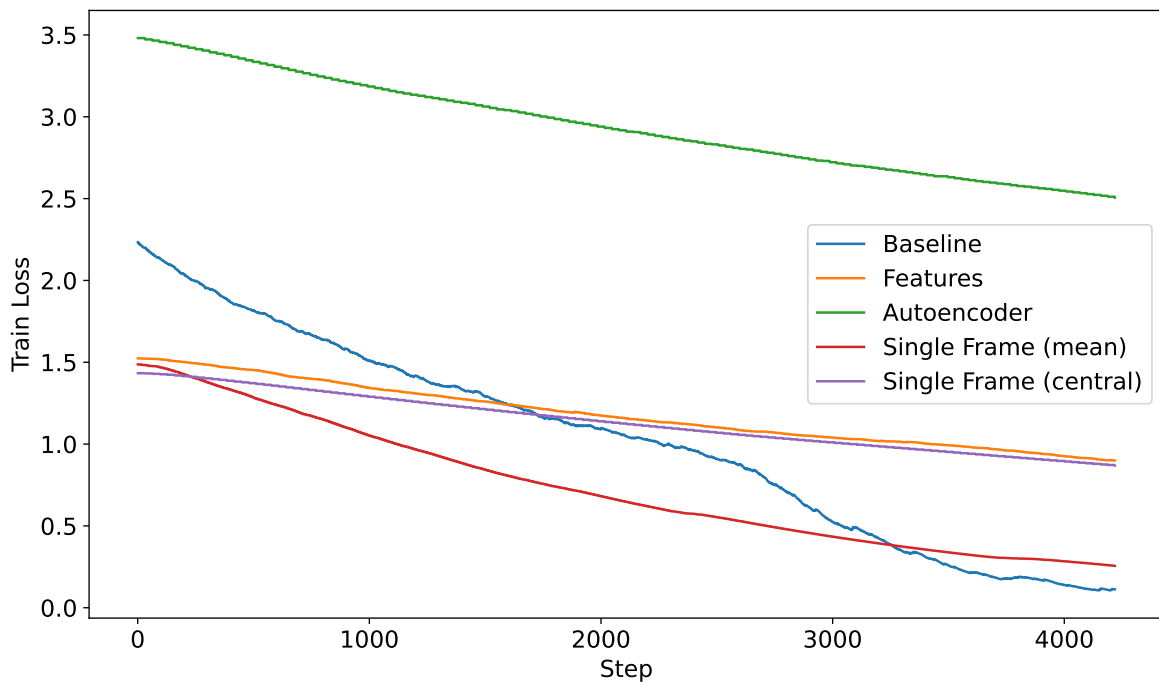


Figure 5: Loss curves over the training procedure for each of the proposed methods.

## 5 Discussion

In this work, we proposed using various dimensionality reduction techniques to classify bird behaviors through videos. To evaluate the performance of the resulting representations, we performed an ablation study, where the most relevant methods were identified. In the following subsections, we will analyze the ablation carried out with the different features and parameters tested from the different dimensionality reduction methods we proposed.

### 5.1 Features

The first comparison we made was between the dataset’s baselines and the method labeled *Features*, where we proposed the use of the architectural backbones of the baseline to extract some features that would later be classified using an MLP. In the comparison shown in fig. 6, we can observe that while the best architecture for the baseline was ResNet3D, the backbone that obtained the most meaningful features was the Swin Transformer.

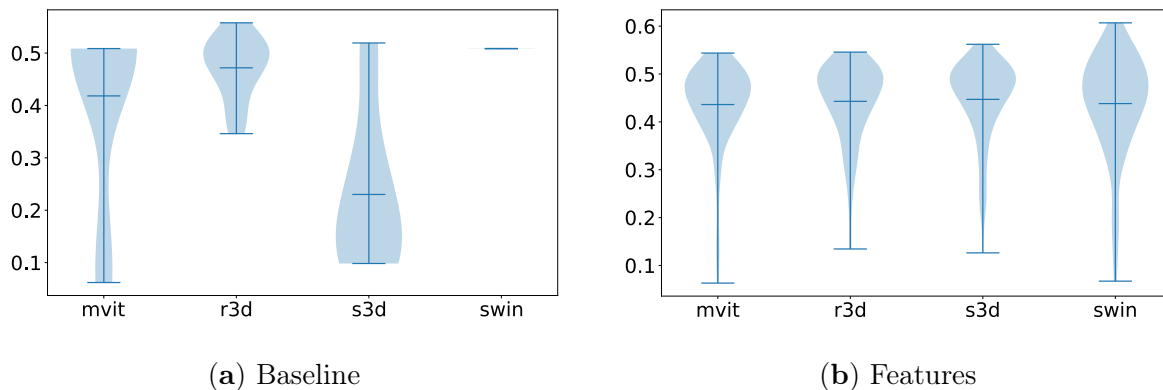


Figure 6: Violin plots illustrating the distribution of parameter importance, comparing the dimensionality reduction method labeled *Features* with the overall performance of the other methods on the Visual WetlandBirds Dataset test set.

### 5.2 Autoencoder

For the *Autoencoder* technique, in fig. 7, we can observe a correlation between certain architecture components and their performance. The information that can be extracted from these plots is that the smaller the encoder, the better performance. As the number of samples in the dataset that we used was not too high, smaller models can better capture insights from the data without massive amounts of overfitting, which hinders the quality of the extracted features.

However, even with the smaller models, the performance remains the lowest among all the approaches, exhibiting significant overfitting. Unlike the other methods, which benefitted from pre-trained weights, training the Autoencoder from scratch on this limited dataset led to its failure in adequately capturing the distribution of diverse actions.

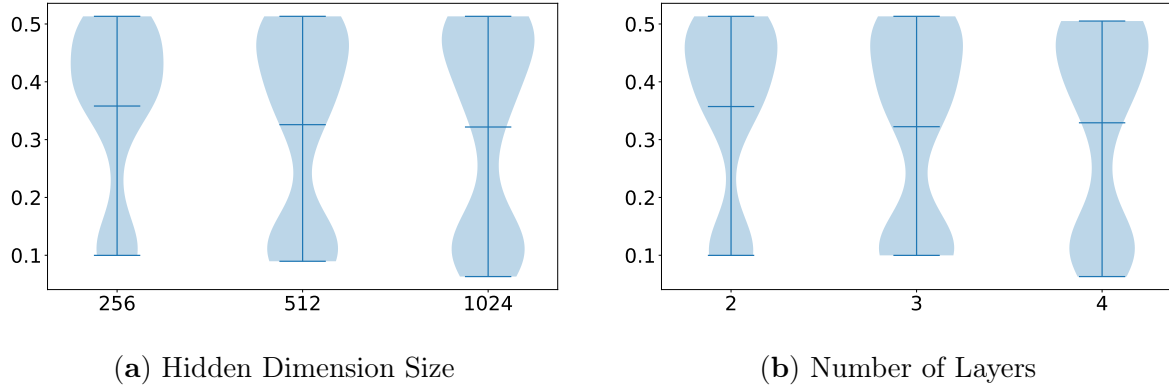


Figure 7: Violin plots illustrating the distribution of parameter importance, comparing the *Autoencoder* dimensionality reduction method with the overall performance of the other methods on the Visual WetlandBirds Dataset test set.

### 5.3 Single Frame

For the *single-frame* technique, fig. 8 illustrates the impact of the method used to select the single frame and the choice of image model on feature extraction. As shown, the mean image provides more information and achieves slightly better performance. Among the image feature extractors, the ResNet architecture produced the best accuracy on the test set.

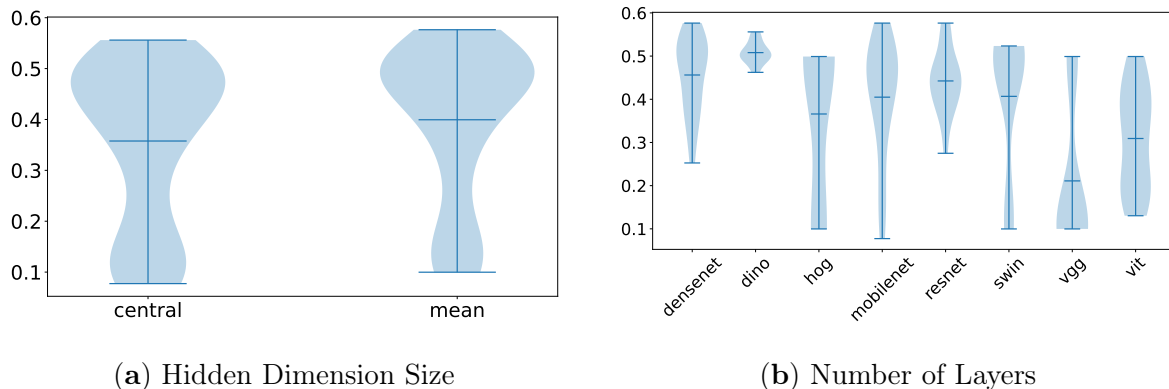


Figure 8: Violin plots illustrating the distribution of parameter importance, comparing the *single-frame* dimensionality reduction method with the overall performance of the other methods on the Visual WetlandBirds Dataset test set.

Although temporal information was lost, this approach demonstrates strong performance, even achieving the best results in certain metrics. Furthermore, while the reduction in data size was not the most substantial, the computation of the reduced representation was significantly faster, as it relied on a single image rather than processing an entire video sequence.

### 5.4 Other Parameters

Finally, figs. 9 to 11 display the effect of different parameters on the performance of the different dimensionality reduction techniques that we proposed. Concerning the batch

size, we can observe that the differences are minimal, with the smallest being slightly better overall. Choosing not to assign class weights to the losses during training led to higher performance across all the methods. And, concerning learning rates, the smallest ones seem to work better with this particular combination of methods and dataset.

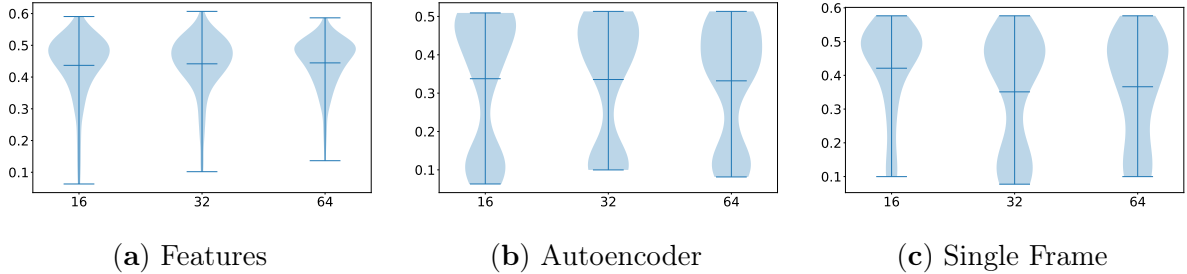


Figure 9: Impact of batch size on the performance of each proposed method.

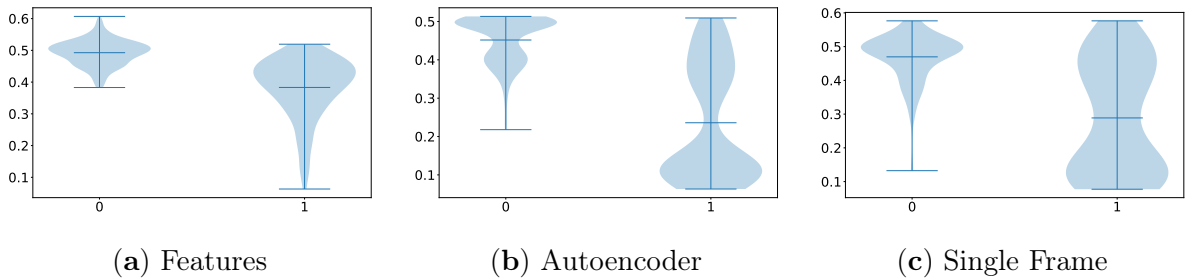


Figure 10: Impact of class weight on the performance of each proposed method.

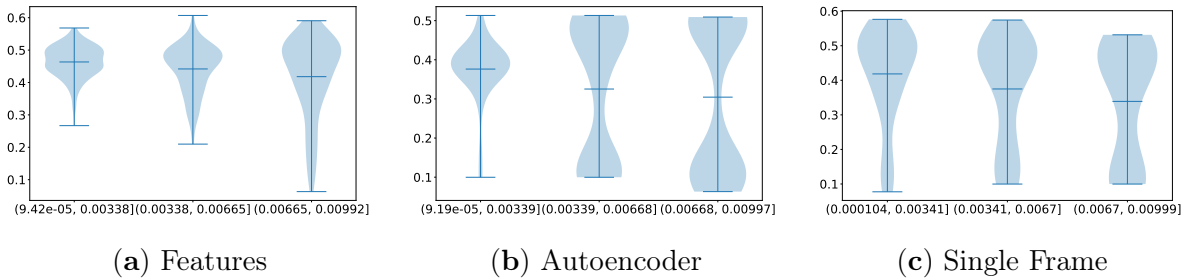


Figure 11: Impact rate on the performance of each proposed method.

## 6 Conclusions

We explored various methods to reduce the dimensionality of videos from the Visual WetlandBirds Dataset for bird action classification. Among the approaches we proposed, the most effective in terms of accuracy and dimensionality reduction utilizes a feature extractor based on a Swing Transformer pre-trained on the Kinetics 400 dataset. This method outperformed the current state-of-the-art techniques while achieving a remarkable reduction in the classifier input size of over 6000 times. As a result, it significantly decreased both the training time and the computational resources required.

While our study was specifically tailored to the Visual WetlandBirds Dataset, the principles we introduced can be applied to other datasets featuring similar challenges—namely, fast movements in noisy natural environments. This flexibility makes our approach well suited for a wide range of animal action videos captured in real-world settings. Consequently, our work opens up promising avenues for future research, especially as new comparable datasets emerge and bring forth fresh challenges in action detection and classification.

Although we conducted an exhaustive ablation study of the methods we tested, several topics for future work remain open. While we implemented various dimensionality reduction techniques, there are other traditional methods we have yet to explore. Additionally, combining the reduced representations we proposed could further enhance the classifiers' performance. Finally, even though it is beyond the scope of this paper, applying these reduced representations to the entire pipeline of action detection and classification could offer valuable insights into the quality of the reduction techniques that we proposed.

## Funding

We would like to thank “A way of making Europe” European Regional Development Fund (ERDF) and MCIN/AEI/10.13039/501100011033 for supporting this work under the “CHAN-TWIN” project (grant TED2021-130890B-C21. HORIZON-MSCA-2021-SE-0 action number: 101086387, REMARKABLE, Rural Environmental Monitoring via ultra wide-Area networkS And distriButed federated Learning). This work is part of the HELEADE project (TSI-100121-2024-24), funded by the Spanish Ministry of Digital Processing and by the European Union NextGeneration EU.

## Data Availability

The dataset used during this study is publicly available at <https://doi.org/10.5281/zenodo.14355257> (accessed on 15 December 2024).

## Acknowledgments

This work was also supported by three Spanish national and two regional grants for PhD studies, FPU21/00414, FPU22/04200, FPU23/00532, CIACIF/2021/430 and CIACIF/2022/175.

## Conflicts Of Interest

The authors declare no conflicts of interest.

## References

- [1] T. O’Riordan, *Environmental Science for Environmental Management*. Longman, 1995. [Online]. Available: <https://books.google.co.uk/books?id=KVxfQgAACAAJ>

- [2] J. D. Nichols and B. K. Williams, “Monitoring for conservation,” *Trends in ecology & evolution*, vol. 21, no. 12, pp. 668–673, 2006.
- [3] C. Margules and M. Usher, “Criteria used in assessing wildlife conservation potential: a review,” *Biological conservation*, vol. 21, no. 2, pp. 79–109, 1981.
- [4] K. S. Smallwood, J. Beyea, and M. L. Morrison, “Using the best scientific data for endangered species conservation,” *Environmental Management*, vol. 24, pp. 421–435, 1999.
- [5] S. Fraixedas, A. Lindén, M. Piha, M. Cabeza, R. Gregory, and A. Lehtikainen, “A state-of-the-art review on birds as indicators of biodiversity: Advances, challenges, and future directions,” *Ecological Indicators*, vol. 118, p. 106728, 2020.
- [6] R. Bellman, R. Bellman, and R. Corporation, *Dynamic Programming*, ser. Rand Corporation research study. Princeton University Press, 1957. [Online]. Available: <https://books.google.es/books?id=rZW4ugAACAAJ>
- [7] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, “A closer look at spatiotemporal convolutions for action recognition,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.
- [8] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, “Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 305–321.
- [9] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, “Video swin transformer,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.
- [10] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, “Mvitv2: Improved multiscale vision transformers for classification and detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4804–4814.
- [11] J. Rodríguez-Juan, D. Ortiz-Perez, M. Benavent-Lledo, D. Mulero-Pérez, P. Ruiz-Ponce, A. Orihuela-Torres, J. García-Rodríguez, and E. Sebastián-González, “Visual wetlandbirds dataset: Bird species identification and behavior recognition in videos,” 2025. [Online]. Available: <https://arxiv.org/abs/2501.08931>
- [12] C. Guo and D. Wu, “Feature Dimensionality Reduction for Video Affect Classification: A Comparative Study,” in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*. Beijing: IEEE, may 2018, pp. 1–6. [Online]. Available: <https://ieeexplore.ieee.org/document/8470329/>
- [13] S. Ayesha, M. K. Hanif, and R. Talib, “Overview and comparative study of dimensionality reduction techniques for high dimensional data,” *Information Fusion*, vol. 59, pp. 44–58, jul 2020. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S156625351930377X>

- [14] G. T. Reddy, M. P. K. Reddy, K. Lakshmana, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker, "Analysis of Dimensionality Reduction Techniques on Big Data," *IEEE Access*, vol. 8, pp. 54 776–54 788, 2020. [Online]. Available: <https://ieeexplore.ieee.org/document/9036908/>
- [15] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction," *Journal of Applied Science and Technology Trends*, vol. 1, no. 1, pp. 56–70, may 2020. [Online]. Available: <https://jastt.org/index.php/jasttpath/article/view/24>
- [16] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559–572, 1901.
- [17] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.
- [18] J. Cohen, "Applied multiple regression," *Correlation Analysis for the Behavioral Sciences/Hillsdale*, 1983.
- [19] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.
- [20] V. Klema and A. Laub, "The singular value decomposition: Its computation and some applications," *IEEE Transactions on Automatic Control*, vol. 25, no. 2, pp. 164–176, 1980.
- [21] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998.
- [22] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
- [23] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
- [24] S. Sra and I. Dhillon, "Generalized nonnegative matrix approximations with bregman divergences," *Advances in neural information processing systems*, vol. 18, 2005.
- [25] G. Chao, Y. Luo, and W. Ding, "Recent Advances in Supervised Dimension Reduction: A Survey," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 341–358, jan 2019. [Online]. Available: <https://www.mdpi.com/2504-4990/1/1/20>
- [26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [27] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

- [28] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition, ed. de rumelhart and j. mcelelland. vol. 1. 1986," *Biometrika*, vol. 71, no. 599-607, p. 6, 1986.
- [29] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [30] G. Dong, G. Liao, H. Liu, and G. Kuang, "A review of the autoencoder and its variants: A comparative perspective from target recognition in synthetic-aperture radar images," *IEEE Geoscience and Remote Sensing Magazine*, vol. 6, no. 3, pp. 44–68, 2018.
- [31] P. Li, Y. Pei, and J. Li, "A comprehensive survey on design and application of autoencoder in deep learning," *Applied Soft Computing*, vol. 138, p. 110176, may 2023. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1568494623001941>
- [32] G. E. Hinton, "Deep belief networks," *Scholarpedia*, vol. 4, no. 5, p. 5947, 2009.
- [33] Y. Kiarashinejad, S. Abdollahramezani, and A. Adibi, "Deep learning approach based on dimensionality reduction for designing electromagnetic nanostructures," *npj Computational Materials*, vol. 6, no. 1, p. 12, feb 2020. [Online]. Available: <https://www.nature.com/articles/s41524-020-0276-y>
- [34] G. Sun, C. Wang, Z. Zhang, J. Deng, S. Zafeiriou, and Y. Hua, "Spatio-temporal Prompting Network for Robust Video Feature Extraction," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, oct 2023, pp. 13 541–13 551. [Online]. Available: <https://ieeexplore.ieee.org/document/10377290/>
- [35] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, "Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 352–364, feb 2018, arXiv:1502.07209 [cs]. [Online]. Available: <http://arxiv.org/abs/1502.07209>
- [36] M. Nottebaum, S. Roth, and S. Schaub-Meyer, "Efficient Feature Extraction for High-resolution Video Frame Interpolation," nov 2022, arXiv:2211.14005 [cs]. [Online]. Available: <http://arxiv.org/abs/2211.14005>
- [37] Z. Lan, Y. Zhu, A. G. Hauptmann, and S. Newsam, "Deep local video feature for action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 1–7.
- [38] H. Ali, M. Sharif, M. Yasmin, M. H. Rehmani, and F. Riaz, "A survey of feature extraction and fusion of deep learning for detection of abnormalities in video endoscopy of gastrointestinal-tract," *Artificial Intelligence Review*, vol. 53, no. 4, pp. 2635–2707, apr 2020. [Online]. Available: <http://link.springer.com/10.1007/s10462-019-09743-2>

- [39] S. H. Abdhussain, A. Rahman Ramli, B. M. Mahmmud, M. Iqbal Saripan, S. Al-Haddad, T. Baker, W. N. Flayyih, and W. A. Jassim, “A Fast Feature Extraction Algorithm for Image and Video Processing,” in *2019 International Joint Conference on Neural Networks (IJCNN)*. Budapest, Hungary: IEEE, jul 2019, pp. 1–8. [Online]. Available: <https://ieeexplore.ieee.org/document/8851750/>
- [40] Yanwei Pang, He Yan, Yuan Yuan, and Kongqiao Wang, “Robust CoHOG Feature Extraction in Human-Centered Image/Video Management System,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 458–468, apr 2012. [Online]. Available: <http://ieeexplore.ieee.org/document/6084765/>
- [41] T. T. Zin, I. Kobayashi, P. Tin, and H. Hama, “A General Video Surveillance Framework for Animal Behavior Analysis,” in *2016 Third International Conference on Computing Measurement Control and Sensor Network (CMCSN)*. Matsue, Japan: IEEE, may 2016, pp. 130–133. [Online]. Available: <http://ieeexplore.ieee.org/document/8008657/>
- [42] J. Fan, N. Jiang, and Y. Wu, “Automatic video-based analysis of animal behaviors,” in *2010 IEEE International Conference on Image Processing*. Hong Kong, Hong Kong: IEEE, sep 2010, pp. 1513–1516. [Online]. Available: <http://ieeexplore.ieee.org/document/5652495/>
- [43] U. Stern, R. He, and C.-H. Yang, “Analyzing animal behavior via classifying each video frame using convolutional neural networks,” *Scientific Reports*, vol. 5, no. 1, p. 14351, sep 2015. [Online]. Available: <https://www.nature.com/articles/srep14351>
- [44] C. Can, X. Yan, and Y. Baoping, “Morphology classification and behaviors identification of birds in scientific video,” in *3rd International Conference on Multimedia Technology (ICMT-13)*. Guangzhou, China: Atlantis Press, 2013, pp. 1449–1457. [Online]. Available: <https://www.atlantis-press.com/article/10530>
- [45] C.-W. Lin, Z. Chen, and M. Lin, “Video-based bird posture recognition using dual feature-rates deep fusion convolutional neural network,” *Ecological Indicators*, vol. 141, p. 109141, aug 2022. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1470160X22006136>
- [46] T. Saito, A. Kanazaki, and T. Harada, “IBC127: Video dataset for fine-grained bird classification,” in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. Seattle, WA, USA: IEEE, jul 2016, pp. 1–6. [Online]. Available: <http://ieeexplore.ieee.org/document/7552915/>
- [47] J. Atanburi, W. Duan, E. Shaw, K. Appiah, and P. Dickinson, “Classification of bird species from video using appearance and motion features,” *Ecological Informatics*, vol. 48, pp. 12–23, nov 2018. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1574954118300566>
- [48] Z. Ge, C. McCool, C. Sanderson, P. Wang, L. Liu, I. Reid, and P. Corke, “Exploiting temporal information for DCNN-based fine-grained object classification,” in *International Conference on Digital Image Computing: Techniques and Applications*, 2016.

- [49] X. L. Ng, K. E. Ong, Q. Zheng, Y. Ni, S. Y. Yeo, and J. Liu, “Animal kingdom: A large and diverse dataset for animal behavior understanding,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 19 023–19 034.
- [50] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The caltech-ucsd birds-200-2011 dataset,” aug 2023.
- [51] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie, “Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 595–604.
- [52] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: <http://arxiv.org/abs/1810.04805>
- [53] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, “Llama: Open and efficient foundation language models,” 2023. [Online]. Available: <https://arxiv.org/abs/2302.13971>
- [54] G. Team *et al.*, “Gemini: A family of highly capable multimodal models,” 2024. [Online]. Available: <https://arxiv.org/abs/2312.11805>
- [55] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, “The kinetics human action video dataset,” 2017. [Online]. Available: <https://arxiv.org/abs/1705.06950>
- [56] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.
- [57] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund, and A. Clapes, “Video Transformers: A Survey ,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 45, no. 11, pp. 12 922–12 943, nov 2023. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/TPAMI.2023.3243465>
- [58] M. Tschannen, O. Bachem, and M. Lucic, “Recent advances in autoencoder-based representation learning,” *arXiv preprint arXiv:1812.05069*, 2018.
- [59] Y. Zhang, “A better autoencoder for image: Convolutional autoencoder,” in *ICONIP17-DCEC*. Available online: [http://users.cecs.anu.edu.au/Tom.Gedeon/conf/ABCs2018/paper/ABCs2018\\_paper\\_58.pdf](http://users.cecs.anu.edu.au/Tom.Gedeon/conf/ABCs2018/paper/ABCs2018_paper_58.pdf) (accessed on 23 March 2017), 2018.
- [60] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, “Dinov2: Learning robust visual features without supervision,” 2023.

- [61] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, “Vision transformers need registers,” 2023.
- [62] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*, vol. 1. Ieee, 2005, pp. 886–893.
- [63] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: <http://arxiv.org/abs/1512.03385>
- [64] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *CoRR*, vol. abs/1704.04861, 2017. [Online]. Available: <http://arxiv.org/abs/1704.04861>
- [65] G. Huang, Z. Liu, and K. Q. Weinberger, “Densely connected convolutional networks,” *CoRR*, vol. abs/1608.06993, 2016. [Online]. Available: <http://arxiv.org/abs/1608.06993>
- [66] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [67] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: <https://arxiv.org/abs/2010.11929>
- [68] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” *CoRR*, vol. abs/2103.14030, 2021. [Online]. Available: <https://arxiv.org/abs/2103.14030>
- [69] O. Déniz, G. Bueno, J. Salido, and F. De la Torre, “Face recognition using histograms of oriented gradients,” *Pattern recognition letters*, vol. 32, no. 12, pp. 1598–1603, 2011.
- [70] B. Bhattarai, R. Subedi, R. R. Gaire, E. Vazquez, and D. Stoyanov, “Histogram of oriented gradients meet deep learning: A novel multi-task deep network for 2d surgical image semantic segmentation,” *Medical Image Analysis*, vol. 85, p. 102747, 2023.