Pig aggression classification using CNN, Transformers and Recurrent Networks

Junior Silva Souza *, Eduardo Bedin†, Gabriel Toshio Hirokawa Higa †, Newton Loebens § and Hemerson Pistori †
* Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil, INOVISÃO
Email: junior.souza@ufms.br
† Dom Bosco Catholic University, Campo Grande, MS, Brazil, INOVISÃO
Email: {ra865659, pistori}@ucdb.br, gabrieltoshio03@gmail.com
§ Federal University of Pampa, Itaqui, RS, Brazil, INOVISÃO
Email: newtonloebens@gmail.com

***Abstract*—The recognition of behavioral relationships related to aggression in pigs is an important task that is visually observed in the livestock industry. However, this task is laborious and susceptible to errors, which can be reduced through automation by visually classifying videos captured in a controlled environment. Video classification can be automated using techniques from computer vision and artificial intelligence, employing neural network methods. The primary techniques utilized in this study are variants of transformers: STAM, TimeSformer, and ViViT, as well as techniques using convolutions, such as ResNet3D2, ResNet(2+1)D, and CnnLstm. These techniques were compared to analyze their individual contributions. The performance was evaluated using metrics such as accuracy, precision, and recall. The TimeSformer technique demonstrated the most promising results in video classification, achieving a median accuracy of 0.729.**

***Index Terms*—Aggressiveness, Video Classification, Transformers, Convolutional.**

I. Introduction

The increasing worldwide demand for protein in the last years has led pig farms to expand. This expansion has been financed by companies that aim at large-scale production to supply the market. In the context of livestock farming, abnormal animal behavior stemming from stress factors due to the environment is one factor that can indicate issues and must be noticed in order to maintain high production levels [1]. Some of the behaviors that require attention are aggressiveness, feeding and mounting. Animal behavior monitoring is usually done manually and requires not only technical knowledge but also experience. It is, therefore, a highly error-prone task [2]. Given the many difficulties involved in this task, the application of technological solutions to monitor animal behavior has been increasing in recent years [3].

The development of deep learning-based computer vision techniques in the last decade has opened new possibilities for automating tasks related to many kinds of problems, including classification, detection, and object tracking. Most relevantly, of course, it has also opened new possibilities for the development of applications aimed at, for example, tracking animals in groups, such as in the work of [4]. In swine farming, cameras can be used to capture images or videos, which can then be processed in order to identify abnormalities [5].

There are important differences between images and videos, as regards the data and information captured. By using videos, it is possible to obtain information about events that occur over a period of time (as opposed to events occurring in a given instant), which can indicate desirable or undesirable behavioral characteristics in swines that occur in short time span. Among the undesirable behaviors, aggressiveness is one that cannot be easily identified by image analysis alone. In such cases, analyzing videos can yield better results by providing temporal information [6].

The fundamental idea of this paper, therefore, is that aggressive behavior can be identified using both spatial information (extracted from individual frames) and temporal information (extracted from sequences of frames), that is, aggressive behavior can be classified through automated video analysis with neural networks. With this hypothesis in mind, this work evaluates video classification techniques applied to the task. More specifically, it evaluates models based on transformers for video processing: Visual Vision Transformer (ViViT), Space Time Attention Model (STAM) and TimeSformer, along with a Convolutional Neural Network (CNN) with Recurrent Neural Network (RNN), ResNet3D and Resnet(2+1)D.

The contribution of this work lies in the creation of a dataset containing videos depicting both aggressive and non-aggressive behaviors, collected within a local commercial breeding environment. Additionally, it proposes the automated classification of aggressive behavior, specifically those occurring in commercial breeding contexts, utilizing computer vision techniques such as convolution, recurrence and transformers for video classification. In this manner, a comparative analysis of the performance of these techniques is conducted, with the objective of determining which among them yields superior results.

Identify applicable funding agency here. If none, delete this.

II. Related Works

The application of deep learning techniques to solve computer vision problems has allowed the development of high-performance techniques for tasks such as image and video classification, semantic segmentation and object detection. The application of these new technologies to animal monitoring has already been the subject of several studies. Some works, such as those by [7] and [8] have noticed the possibility of using images of pig faces to monitor variables such as welfare and mood, and also as an auxiliary management tool in general. Going beyond their field, computer vision techniques have also influenced research such as that by [9], which proposes the TransformerCNN model to monitor pigs by processing their vocal sounds. [10] studied the application of feature pyramid networks (FPNs) to segment individual pigs. When behavior is specifically concerned, there are also important works. [11] proposed a system to recognize feeding behavior using object detection networks, such as YoloV3. [12] proposed an algorithm to detect drinking, urination and mounting behavior in real-time, by processing images generated in a pig pen. In these cases, the primary focus was on still images. Videos were used as a way of producing singular images more efficiently, which is a common practice since the cameras can be left hanging over the pig pen, or attached to something, and configured to take videos from which individual frames can be sampled as pictures.

On the other hand, [13] used convolutional neural networks to process videos, in order to identify five behaviors of interest: sleeping, mounting, lying down, feeding and walking. The proposed approach used optical flow techniques to extract movement information, and convolutions were applied in parallel on RGB video frames. The detection and monitoring of animals were also analyzed by [14], who performed identification and monitoring by using CNN and Kalman filter for tracking and analyzing boars afflicted with African swine fever. In this last case, the deep learning techniques were used as part of another research aiming at proving the point that this sickness can be identified early by monitoring the movement of the animals. Another relevant case where deep learning was used within research is the work by [15], where the interaction of animals with enrichment objects was analyzed as a way of reducing aggressive behavior. Finally, [16] applied a CNN with Long Short-Term Memory (LSTM) to classify videos of piglets according to their postures. The authors utilized PCA to select frames, reducing their numbers and facilitating network training. As stated above, many works use videos to capture still images, and process still images for information. For many problems, single images are enough for classification using neural networks. However, these works show that behavior monitoring can benefit from the extraction of deeper spatial and temporal information from sequential image data.

There are recent developments in the field of computer vision that may prove useful for animal behavior monitoring, and have not yet been properly evaluated. Above all, these are represented by the family of transformer-based architectures, whose central component is their attention mechanism. The use of attention in image and video classification allows the implementation of techniques aimed at highlighting important features present in the data, along with their relationship [17], [18]. [19] proposed to apply attention concepts to (still) image processing. Then, [20] proposed to expand their application to video processing. Given the current state of arts, in this work, we present a video dataset with videos of aggressive behaviors of pigs. Furthermore, it includes an evaluation of transformer-based, recurrent, and convolutional neural techniques for video classification, specifically applied to identify instances of aggressive behavior.

This research proposed a comparison of different techniques for video classification, focusing primarily on methods that utilize transforms, convolutions, and recurrent neural networks, as seen in various prior studies. However, such a comparison had not been conducted until this work was undertaken.

III. Materials and Methods

A. Video dataset

For this research, a video dataset was collected in a private property in the state of Mato Grosso do Sul, Brazil. The dataset includes videos of pigs that were raised as livestock and were between 60 to 180 days old at the time of capture. The images were collected from March until July of 2022, when the mean temperature was 28°C. The videos were captured mostly during the early morning hours, between 08h00 to 10h00, and the late afternoon hours, between 13h00 to 16h00, when the animals were more active.

The dataset consists of 421 video clips, of which 143 were annotated as Aggressive and 278 were annotated as Not Aggressive. The videos were captured from a height of 2.6 meters and a distance of 5 meters from each enclosure, utilizing a fixed diagonal camera position. The camera device used was a Motorola e(6i) configured to capture frames with a height of 2340 pixels and a width of 4160 pixels, using the RGB color space at a rate of 30 frames per second, in recording periods of 39 minutes for each video, which were used to obtain the clips. A total of 6 videos were obtained, totaling approximately 240 minutes, of which 40 minutes were used to generate the clips.

To obtain the dataset, it was necessary to edit each video recording[1]. The videos were manually segmented into clips with 80 frames at moments when the animals exhibited behavior considered aggressive, as well as moments when they did not display aggression but were in close proximity. To consider an animal aggressive, it was observed that the animal head made movements towards the body of the other animal, while the latter attempted to evade. In Figure 1 we can observe an example of an input video recording composed of 70,200 frames (39 minutes at 30 frames per second), where the manual classification (aggression or nonaggression) was performed. Furthermore, it is possible to observe that the

[1]The editing process was performed using the Python programming language and the OpenCV package.

image presents a blue tone due to the barn fence on the side of the enclosure where the animals are housed. This is done to provide both light and ventilation during the day, a practice adopted by many local shelters.

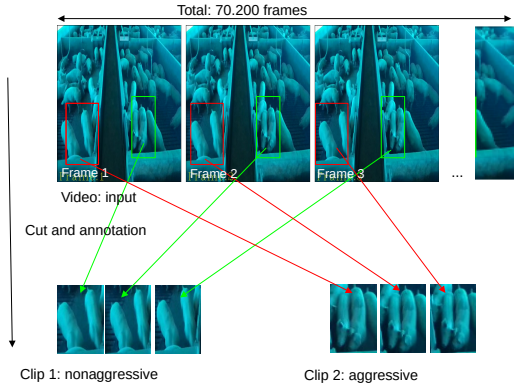

Fig. 1. In this image, we can visualize the process of classification from an input video, where clips are obtained.

In Figure 2, we can see an example of three frames that depict two pigs in an aggressive moment from a clip. The animals involved, pig 1 and pig 2, display an initial aggression, attempting to bite each other. This is a brief moment that occurs within a short time interval.

Fig. 2. The image shows three sequential frames depicting two instances of aggression. We highlighted with white polygon details in frame three, located on the right, where the animals are attempting to bite.

The animal behavior, considered non-aggressive, can be observed in Figure 3, where we can see animals in standing up and animals lying down in normal conditions.

Fig. 3. Image depicting three frames illustrating non-aggressive animal behavior.

B. Video preprocessing

Video compression was employed to reduce the number of frames to be processed by eliminating redundant information from the clips. In the work of [21], the reduction of frames was analyzed to determine the minimum and sufficient amount of frames needed for training and classification. This processing allowed for a reduction of processing during the training. The reduction of clip frames was accomplished at a 4:1 ratio. As a result, each sequence of 40 frames was condensed to 10 frames. Besides the compression, each frame of video was also resized to 400 pixels of height and 400 pixels in width.

C. Deep Learning

This paper compares the performance of deep neural networks applied to video classification, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers-based neural networks. Convolutional neural networks, either combined with or without recurrent neural networks, are considered as alternative approaches for video classification, as demonstrated in previous works such as [22] and [23].

To perform the classification of videos using convolutional neural networks, architectures have been proposed in [24] that utilize the mechanism of "skip connection"to handle the problem of vanishing gradients which can reduce the magnitude of the updates of neurons in training. In order to work with spatial and temporal information, two architectures have been developed: ResNet3D and Resnet(2+1)D also referring to as ResNet21D in this work. ResNet3D performs convolutions in three dimensions to extract both temporal and spatial information from video frames. On the other hand, the ResNet(2+1)D architecture performs 2D convolutions followed by 1D convolutions, extracting spatial and temporal information, respectively.

The use of recurrent neural networks (RNNs) in combination with convolutional neural networks (CNNs) was also proposed in this work. The network used was ResNet. The CNN was utilized for extracting features from videos, which represent spatial information. Meanwhile, the RNN was employed for classification and extraction of information from frames, representing temporal information [25].

Transformers are implementations of attention mechanisms developed initially in the field of natural language processing. They have since been adapted for use in computer vision in tasks such as image and video classification, detection and segmentation [26].

In natural language processing, each word is represented in the form of tokens. However, in computer vision, an image is divided into small parts called patches, which are used as tokens, and are then used for image classification using ViT (Vision Transformers).

The ViT introduces the construction of a neural network that uses transformers, as an adaptation that does not use convolutions [19], [27]. The tokens are extracted from an image and passed to a transformer encoder which corresponds to one block of the transformer. One of the differentials of

the transformer is its attention mechanism, specifically the dot product attention presented in the encoder. The encoders are responsible for extracting information, which is used by a multilayer perceptron that is responsible for classification.

The ViViT architecture is an application of ViT in video and image classification problems. In this architecture, each video is treated as a sequence of images from which important spatial and temporal information are extracted for classification purposes. As videos are composed of sequential images (frames), relevant information can be extracted at each frame change. Furthermore, each frame in the video can be analyzed by ViT, which extracts spatial information that is used to extract temporal information. The architecture ViViT has two stages of application using ViT: The first stage is applied to each frame from the video and the second stage is applied to the results from the first stage [28].

The STAM architecture is similar to ViViT. This architecture applies ViT to extract spatial and temporal information, which gives more efficient processing while prioritizing the extraction of spatial information before capturing temporal information. In STAM, 16 frames are selected per video for processing, as described in [29].

The TimeSformer (Time-Space Transformer) architecture divides each frame of the video into patches with a size of 16x16, which can be changed to other sizes such as 32x32, 64x64. This architecture uses ViT to extract spatial and temporal information from the obtained patches. During processing, only the patches with more information are selected. The quantity and location of patches can vary during the processing, which involves the execution of multiple ViT models. This processing allows the model to focus on the most important parts of each frame to achieve video classification [30].

D. Experimental Setup

To conduct the experiments, the dataset was used with a 5-fold cross-validation [31]. In each run, 20% of the training images were used for validation. The frames of the videos were resized to $(64, 64)$ pixels due to the limitations of the runtime environment, which used the NVIDIA TITAN Xp with 12 GB of memory, supporting execution with sizes no larger than that. For the *transformer* modules, 32 patches were extracted from each frame. This initial value is used in works such as [32]. Each architecture was tested with the following learning rates: 0.0001 and 0.0003, values utilized in different works, such as that of [33], as well as that of [34]. In this work, the neural networks were optimized with Adam, which yielded good results in works such as that of [35] and [36].

The training was performed for 20 epochs, with batches of size 8, following the approach by [37]. The weights were randomly initialized, and the transfer learning technique was not utilized due to the specific characteristics of the application domain. At the end of the training phase, four classification metrics were calculated for the test folds, and the mean values were obtained for the following metrics: accuracy, precision, recall, and F-score. Then, these results were evaluated through an ANOVA hypothesis test, with a significance threshold of 5%. The Scott-Knott clustering test was used post-hoc, when the ANOVA results were significant.

IV. Results

The Table I shows the values of accuracy, precision, recall, and F-score metrics. The table columns display the names of the techniques, as well as the metrics along with their average and Standard Deviation (SD) values. The values highlighted in bold represent the best results in terms of metrics and techniques used, while the letters 'a' and 'b' represent the defined clusters obtained by Scott-Knott.

Techniques	**Metrics**			
	Accuracy (SD)	**Precision (SD)**	**Recall (SD)**	**F-score (SD)**
ViViT	0.698 (0.056) a	0.580 (0.111) a	0.381 (0.129) a	0.454 (0.125) a
STAM	0.665 (0.016) a	0.300 (0.358) b	0.035 (0.046) b	0.062 (0.082) b
TimeSformer	**0.764 (0.083) a**	**0.716 (0.125) a**	0.499 (0.207) a	0.573 (0.178) a
CnnLstm	0.724 (0.092) a	0.614 (0.142) a	0.480 (0.189) a	0.532 (0.168) a
Resnet3D	0.750 (0.107) a	0.670 (0.197) a	0.531 (0.187) a	0.585 (0.185) a
Resnet21D	0.757 (0.111) a	0.653 (0.183) a	**0.613 (0.225) a**	**0.621 (0.193) a**

TABLE I
THIS TABLE DESCRIBES THE AVERAGE VALUES OF ALL METRICS USED FOR VIDEO CLASSIFICATION TECHNIQUES.

The TimeSformer technique presented the best performance in the accuracy and precision metrics, with value 0.764 for their respective means. However, the Resnet21D technique outperformed TimeSformer in the recall metric.

The ANOVA results were marginally significant for the techniques in all four metrics, with p-values of 0.0771, 0.000449, 1.2×10^{-8}, and 4.9×10^{-10} for accuracy, precision, recall, and F-score, respectively. These values were analyzed and indicate statistically significant differences between almost all analyzed techniques, except for accuracy when considering a significance level of 5%. Subsequent post-hoc analysis using the Scott-Knott clustering test revealed that TimesFormer outperformed the other techniques in precision metrics, despite being clustered together with ViViT, Resnet3D, Resnet21D and CnnLstm. STAM achieved the worst performance in all metrics.

The significant differences between the techniques can be analyzed using the precision metric, which yielded a p-value of 0.000009. These values indicate strong statistical differences among the adopted techniques. Furthermore, when comparing the learning rate, the precision and recall metrics also showed significant differences, with p-values of 0.049393 and 0.01480, respectively. Considering the metrics of precision and a learning rate of 0.0001, the technique that showed the best result is TimeSformer, which is a variant of the transformer model.

We can visualize in Table II four video clips, with each clip displaying three frames as an illustrative example. The GT (Ground Truth) column represents the actual outcomes, while the DL (Deep Learning) column reflects the results from TimeSformer classification technique. The clips identified as 'a' was correctly classified as 'NA' (Non-Aggressive), whereas the clips 'b' was accurately classified as 'A' (Aggressive). Additionally, in Table II, we can observe the clips 'c' and 'd',

which, although categorized as 'NA' and 'A', respectively, in the GT column, were not classified correctly.

Frames	GT	DL
(a) 	NA	NA
(b) 	A	A
(c) 	NA	A
(d) 	A	NA

TABLE II

ACCURATE AND INACCURATE CLASSIFICATION OF VIDEOS ALONG WITH THEIR CORRESPONDING LABELS. TAKING INTO ACCOUNT NA (NON-AGGRESSIVE), A (AGGRESSIVE), GT (GROUND TRUTH), AND DL (DEEP LEARNING TECHNIQUES). AN INTERVAL OF 3 FRAMES WAS USED BETWEEN FRAMES IN THE CLIPS.

The confusion matrix was obtained for the TimeSformer technique in the testing dataset, as shown in Figure 4. We can observe an average accuracy rate 0.764 in the classification, where there are more videos classified as non-aggressive.

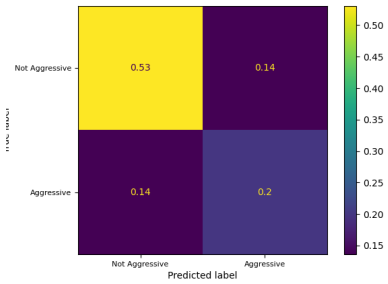

Fig. 4. The confusion matrix obtained from the testing set using the TimeSformer technique.

V. DISCUSSION

In this research, we propose the use of variants of transformers for video classification. The transformer technique has been applied in works involving videos, such as [38], and has shown better results when using datasets with videos of gastrointestinal diseases. However, the use of transformers in computer vision presents a challenge, as transformers rely on extensive datasets to yield improved results, necessitating high-quality images and substantial processing power, as discussed in [39].

The reduction of the number of processed frames is an interesting strategy adopted to improve processing efficiency and reduce redundant information. However, it can affect the model's generalization capacity. In the case of STAM, which was configured to process 16 frames per video, it did not result in better precision, recall and F-score. This may be due to the fact that the behavior of interest is not always clearly presented in the initial frames of the analyzed videos.

In the case of ViViT, which is another variant of transformer-based models, it did not show better results compared by the analyzed metrics. This could be attributed to differences in architecture and the way ViViT processes and extracts information from frames. On the other hand, TimeSformer emerged as a better option, particularly in terms of precision. Its approach of selecting relevant patches from each frame allows for reducing redundant information and focusing on the most important characteristics for classification. This strategy of patch selection proves advantageous for the classification task and leads to superior results, especially in terms of precision.

Although the use of CnnLstm with recurrent neural networks has shown better results in previous works, such as [40], this technique did not demonstrate the best performance in this project. This can be attributed to the fact that the analyzed videos consist of series of frames that require the processing of long sequences, which becomes a bottleneck for an LSTM-based to approach.

However, it is important to emphasize that the other techniques that exclusively used convolution, such as Resnet3D and Resnet21D, achieved better results compared to the CnnLstm approach. This indicates that combining LSTM with convolution does not yield significant improvements in this specific context.

The limitation of this work is the size of database, which needs to be larger. however, capturing and selecting data in real production is challenging and requires time and effort. Therefore, for future work, the goal is to increase the size of database for new experiments, reinforcing the use of techniques that extract both spatial and temporal information. Additionally, the balanced class is another issue to be improved with data augmentation.

VI. CONCLUSION

This work proposed the use of deep learning techniques, including transformers, convolutional neural networks, and recurrent layers to classify aggressive behavior in pigs. To accomplish this, a video dataset was created for training and validating the employed techniques. The techniques were evaluated using accuracy, precision, recall, and F-measure metrics, which demonstrated that the TimeSformer technique achieved the best results in video classification across all metrics. While variants of transformers have been utilized in numerous works involving video classification, the TimeSformer technique outperformed them with the dataset created and used in this study, particularly in terms of precision.

For future work, the exploration of a varied number of frames per clip and the utilization of techniques related to optical flow can be considered to improve the accuracy of aggressive behavior classifications. Moreover, incorporating techniques related to optical flow can improve the detection

of subtle movements in videos, allowing for a more robust analysis of pig behavior. This can be particularly useful for identifying aggressive behavior in the early stages.

REFERENCES

[1] L. Lassaletta, F. Estellés, A. H. Beusen, L. Bouwman, S. Calvet, H. J. Van Grinsven, J. C. Doelman, E. Stehfest, A. Uwizeye, and H. Westhoek, "Future global pig production systems according to the shared socioeconomic pathways," *Science of the Total Environment*, vol. 665, pp. 739–751, 2019.

[2] A. Alameer, I. Kyriazakis, H. A. Dalton, A. L. Miller, and J. Bacardit, "Automatic recognition of feeding and foraging behaviour in pigs using deep learning," *biosystems engineering*, vol. 197, pp. 91–104, 2020.

[3] H. Shao, J. Pu, and J. Mu, "Pig-posture recognition based on computer vision: Dataset and exploration," *Animals*, vol. 11, no. 5, p. 1295, 2021.

[4] J. Xu, S. Zhou, A. Xu, J. Ye, and A. Zhao, "Automatic scoring of postures in grouped pigs using depth image and cnn-svm," *Computers and Electronics in Agriculture*, vol. 194, p. 106746, 2022.

[5] S. Ma, Q. Zhang, T. Li, and H. Song, "Basic motion behavior recognition of single dairy cow based on improved rexnet 3d network," *Computers and Electronics in Agriculture*, vol. 194, p. 106772, 2022.

[6] M. Wang, M. L. Larsen, D. Liu, J. F. Winters, J.-L. Rault, and T. Norton, "Towards re-identification for long-term tracking of group housed pigs," *Biosystems Engineering*, vol. 222, pp. 71–81, 2022.

[7] M. F. Hansen, E. M. Baxter, K. M. Rutherford, A. Futro, M. L. Smith, and L. N. Smith, "Towards facial expression recognition for on-farm welfare assessment in pigs," *Agriculture*, vol. 11, no. 9, p. 847, 2021.

[8] W. Shigang, W. Jian, C. Meimei, and W. Jinyang, "A pig face recognition method for distinguishing features," in *2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*. IEEE, 2021, pp. 972–976.

[9] J. Liao, H. Li, A. Feng, X. Wu, Y. Luo, X. Duan, M. Ni, and J. Li, "Domestic pig sound classification based on transformercnn," *Applied Intelligence*, pp. 1–17, 2022.

[10] Z. Hu, H. Yang, and T. Lou, "Dual attention-guided feature pyramid network for instance segmentation of group pigs," *Computers and Electronics in Agriculture*, vol. 186, p. 106140, 2021.

[11] M. Kim, Y. Choi, J.-n. Lee, S. Sa, and H.-c. Cho, "A deep learning-based approach for feeding behavior recognition of weanling pigs," *Journal of Animal Science and Technology*, vol. 63, no. 6, pp. 1453–1463, 2021. [Online]. Available: https://doi.org/10.5187/jast.2021.e127

[12] Y. Zhang, J. Cai, D. Xiao, Z. Li, and B. Xiong, "Real-time sow behavior detection based on deep learning," *Computers and Electronics in Agriculture*, vol. 163, p. 104884, 2019.

[13] K. Zhang, D. Li, J. Huang, and Y. Chen, "Automated video behavior recognition of pigs using two-stream convolutional networks," *Sensors*, vol. 20, no. 4, p. 1085, 2020.

[14] E. Fernández-Carrión, J. Á. Barasona, Á. Sánchez, C. Jurado, E. Cadenas-Fernández, and J. M. Sánchez-Vizcaíno, "Computer vision applied to detect lethargy through animal motion monitoring: a trial on african swine fever in wild boar," *Animals*, vol. 10, no. 12, p. 2241, 2020.

[15] C. Chen, W. Zhu, M. Oczak, K. Maschat, J. Baumgartner, M. L. V. Larsen, and T. Norton, "A computer vision approach for recognition of the engagement of pigs with different enrichment objects," *Computers and Electronics in Agriculture*, vol. 175, p. 105580, 2020.

[16] M. Wang, M. Larsen, F. Bayer, K. Maschat, J. Baumgartner, J.-L. Rault, T. Norton *et al.*, "A pca-based frame selection method for applying cnn and lstm to classify postural behaviour in sows," *Computers and Electronics in Agriculture*, vol. 189, p. 106351, 2021.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] X. Dai, Y. Chen, B. Xiao, D. Chen, M. Liu, L. Yuan, and L. Zhang, "Dynamic head: Unifying object detection heads with attentions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7373–7382.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[20] 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," 2021.

[21] D. Liu, M. Oczak, K. Maschat, J. Baumgartner, B. Pletzer, D. He, and T. Norton, "A computer vision-based method for spatial-temporal action recognition of tail-biting behaviour in group-housed pigs," *Biosystems Engineering*, vol. 195, pp. 27–41, 2020.

[22] P. Patil, V. Pawar, Y. Pawar, and S. Pisal, "Video content classification using deep learning," *arXiv preprint arXiv:2111.13813*, 2021.

[23] D. K. Singh, M. A. Ansari, and S. Pallawi, "Computer vision based visual activity classification through deep learning approaches," in *2022 IEEE Region 10 Symposium (TENSYMP)*, 2022, pp. 1–5.

[24] 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.

[25] C. Harris, K. R. Finn, M.-L. Kieseler, M. R. Maechler, and P. U. Tse, "Deepaction: a matlab toolbox for automated classification of animal behavior in video," *Scientific Reports*, vol. 13, no. 1, p. 2688, 2023.

[26] J. Bi, Z. Zhu, and Q. Meng, "Transformer in computer vision," in *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, 2021, pp. 178–188.

[27] Z. Fu, "Vision transformer: Vit and its derivatives," 2022.

[28] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 6836–6846.

[29] G. Sharir, A. Noy, and L. Zelnik-Manor, "An image is worth 16x16 words, what is a video worth?" 2021.

[30] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" 2021. [Online]. Available: https://arxiv.org/abs/2102.05095

[31] W. Liu, W. Luo, Z. Li, P. Zhao, S. Gao *et al.*, "Margin learning embedded prediction for video anomaly detection with a few anomalies." in *IJCAI*, 2019, pp. 3023–3030.

[32] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 7262–7272.

[33] M. M. Islam, N. Tasnim, and J.-H. Baek, "Human gender classification using transfer learning via pareto frontier cnn networks," *Inventions*, vol. 5, no. 2, p. 16, 2020.

[34] A. Nasirahmadi, B. Sturm, S. Edwards, K.-H. Jeppsson, A.-C. Olsson, S. Müller, and O. Hensel, "Deep learning and machine vision approaches for posture detection of individual pigs," *Sensors*, vol. 19, no. 17, 2019.

[35] C. Desai, "Comparative analysis of optimizers in deep neural networks," *International Journal of Innovative Science and Research Technology*, vol. 5, no. 10, pp. 959–962, 2020.

[36] A. Almadani, A. Shivdeo, E. Agu, and J. Kpodonu, "Deep video action recognition models for assessing cardiac function from echocardiograms," in *2022 IEEE International Conference on Big Data (Big Data)*, 2022, pp. 5189–5199.

[37] W. Zhou, Y. Zhu, J. Lei, R. Yang, and L. Yu, "Lsnet: Lightweight spatial boosting network for detecting salient objects in rgb-thermal images," *IEEE Transactions on Image Processing*, vol. 32, pp. 1329–1340, 2023.

[38] C. Zhang, A. Ding, Z. Fu, J. Ni, Q. Chen, Z. Xiong, B. Liu, Y. Cao, S. Chen, and X. Liu, "Deep learning for gastric location classification: An analysis of location boundaries and improvements through attention and contrastive learning," *Smart Health*, vol. 28, p. 100394, 2023.

[39] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys*, vol. 54, no. 10s, pp. 1–41, 2022.

[40] J. Han, J. Siegford, D. Colbry, R. Lesiyon, A. Bosgraaf, C. Chen, T. Norton, and J. P. Steibel, "Evaluation of computer vision for detecting agonistic behavior of pigs in a single-space feeding stall through blocked cross-validation strategies," *Computers and Electronics in Agriculture*, vol. 204, p. 107520, 2023.