Exploring Contrastive Learning for Video Classification on Harmful Content

Tiancheng Qin
Faculty of Engineering
University of Ottawa
Ottawa, Canada
tqin021@uottawa.ca

Kelvin Mock
Faculty of Engineering
University of Ottawa
Ottawa, Canada
kmock073@uOttawa.ca

***Abstract*—Harmful content on video sharing platforms poses significant risks to viewers, especially minors. Detecting such content is crucial to prevent its spread and protect users. In this paper, we explore the impact of contrastive learning on the classification of harmful content in videos. We compared three methods: a baseline without contrastive learning, a multi-dataset supervised contrastive learning approach (MSupCL), and a self-supervised contrastive learning approach (SSCL). We used two feature extractors, C3D and R(2+1)D_18, to evaluate how contrastive learning affects different architectures. Our experiments on the Real-Life Violence Videos dataset and the TikHarm dataset showed that contrastive learning does not significantly improve classification performance in this setting. We will discuss the challenges encountered and suggest directions for future research. The code can be viewed on Github: https://github.com/kmock930/Harmful-Video-Contrastive-Classification**

***Index Terms*—Harmful Content Classification, Contrastive Learning, Video Detection, Feature Extractors, Deep Learning**

I. Introduction

A. Background

There is a significant risk to viewers, especially minors, from harmful content on video sharing platforms. In their survey, A. Arora et al. [1] emphasized the importance of detecting harmful content on social platforms (e.g., TikTok), noting that such platforms face significant challenges in reducing the spread of harmful content while ensuring a safer user experience. Because human filtering is difficult to scale, researchers are working to develop automated tools to help with content review, but there is a significant gap between current research and platform needs.

Video data contains both temporal and spatial information, and it is a very complex process to extract useful features from it to distinguish between violent and non-violent behaviors. Unlike static images, the dynamic feature in video requires that the model not only learns the image content but also be able to understand the motion between consecutive frames, which makes feature representation very difficult. [2]

Contrastive learning offers a way to learn representations from data pairs by bringing similar instances closer and pushing dissimilar ones apart in the feature space. [3] This reduces the reliance on labeled data and can improve model generalization.

B. Objectives

In this study, we investigated the impact of contrastive learning on harmful content video classification. By conducting a transfer learning, we adapted the multi-dataset supervised contrastive learning (MSupCL) method [3], originally used for early autism detection, to the domain of harmful video detection. [3] We also compared it with the self-supervised contrastive learning method SSCL (also called SimCLR) [4].

C. Contributions

We used two feature extractors, C3D and R(2+1)D_18 [5], [6], to evaluate the effect of contrastive learning on different architectures. Our experiments were carried out on two datasets: the Real-Life Violence Videos dataset and the TikHarm dataset. We will analyze the results and discuss the challenges encountered.

II. Related Work

Video classification has been extensively studied using deep learning techniques. Convolutional Neural Networks (CNNs) are commonly used for extracting spatial features, while Recurrent Neural Networks (RNNs) capture temporal dynamics. [7]

A. Video Processing Network

The C3D model proposed by D. Tran et al. [5] generates a compact feature representation of a video by using a $3 \times 3 \times 3$ convolutional kernel to capture both spatial and temporal information in the video data. Its contribution lies in its efficiency and simplicity for a variety of video analysis tasks, including action recognition and scene classification. And the author later proposed "(2+1)D" convolution [6], which decomposes the 3D convolution into independent 2D spatial convolution and 1D temporal convolution. This method improves the nonlinear representation of the network by separating spatial and temporal information.

B. Contrastive Learning in Visual Representations

Contrastive learning has gained attention for its ability to learn representations without labeled data. Chen et al. [4] introduced SimCLR, a simple framework for contrastive learning of visual representations. Similar uses data augmentations

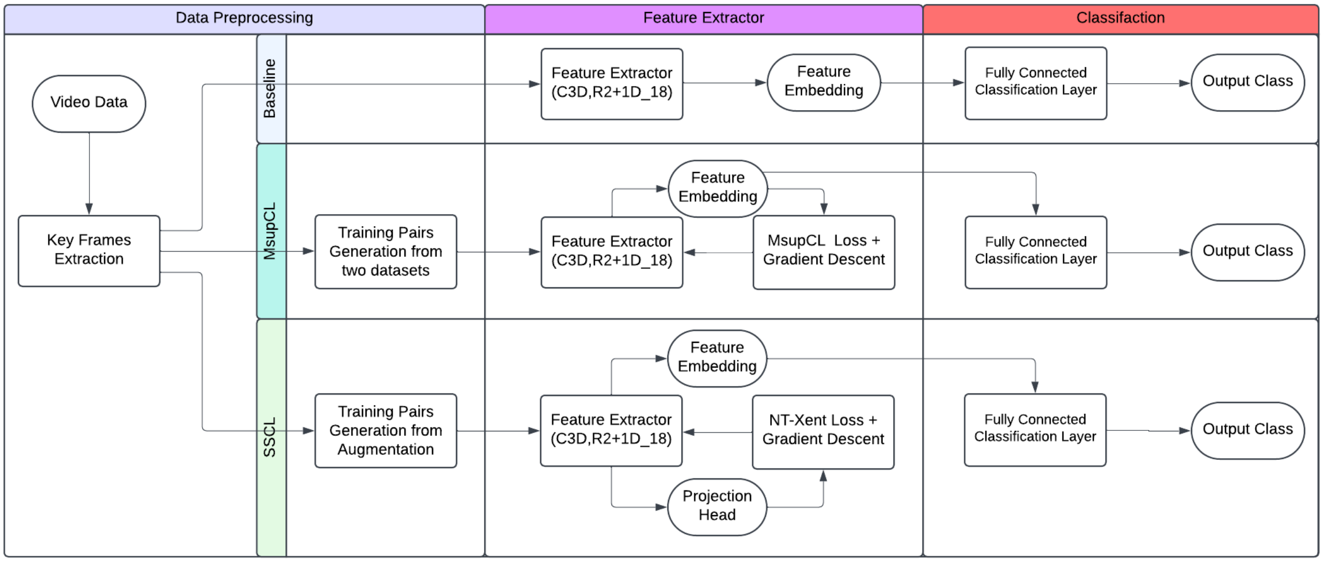

Data Preprocessing
Feature Extractor
Classifaction
Video Data
Key Frames Extraction
Baseline
MsupCL
SSCL
Training Pairs Generation from two datasets
Training Pairs Generation from Augmentation
Feature Extractor (C3D,R2+1D_18)
Feature Embedding
MsupCL Loss + Gradient Descent
NT-Xent Loss + Gradient Descent
Projection Head
Fully Connected Classification Layer
Output Class

Fig. 1: Complete Architecture of our Experiments.

and a contrastive loss function to learn features suitable for downstream tasks. Rani and Verma [3] proposed the MSupCL method for early autism diagnosis. They used a supervised contrastive loss to learn discriminative features from multiple datasets with diverse distributions.

C. *Violence detection in videos*

F. Ullah et al. [7] compared traditional machine learning methods such as SVM, BoW, etc. with deep learning methods such as CNN, LSTM, C3D, R(2+1)D networks in several video violence detection tasks. They concluded that deep learning methods show better performance in this task, but due to their complexity and other issues, it is difficult to deploy them in practical applications. Such deep learning methods are still the focus of future research in the violence detection area.

III. Methodology

Our experimental setup consists of three main components: data preprocessing, feature extraction, and classification. We compared three methods: a baseline without contrastive learning, a multi-dataset supervised contrastive learning (MSupCL) approach [3], and a self-supervised contrastive learning (SSCL) approach [4].

A. *Data Preprocessing*

Data preprocessing depends on the contrastive learning method used. For the baseline method, we process each video individually. For contrastive learning methods, we generated pairs of data for computing the contrastive loss. All the video data were extracted as averaged frames with the same size as the inputs to the feature extractor models. The data for the SSCL method was then applied with data augmentation in order to obtain augmented views that were formed into contrasting pairs.

B. *Feature Extractors*

We used two feature extractors for the frames data to generate 128-dimension embedding:

- C3D Model [5]: Captures spatiotemporal features using 3D convolutions. A $3\times3\times3$ convolution kernel was used here.
- R(2+1)D_18 Model [6]: Decomposes 3D convolutions into separate spatial (2D) and one additional temporal convolutions (1D). An $1 \times 3 \times 3$ convolution kernel was used in the 2D convolutions and a $3 \times 1 \times 1$ convolution kernel was used for 1D convolution.

C. *Classification Network*

The classification network is a fully connected layer with softmax activation. It takes feature embedding output from the feature extractor as an input. Then, it will map the features embeddings to class probabilities. We used the Sparse Categorical Crossentropy loss and the Adam optimizer with an initial learning rate of 1e-5. The learning rate was multiplied by 0.01 every three epochs. We trained for 10 epochs by default. It has been evaluated with the Sparse Categorical Accuracy metric.

D. *Method*

1) Baseline Model (BinClass): We performed feature extraction and classification on individual datasets without training the feature extractor. The features were extracted using feature extractors and passed to the classification network to make the final decision.

2) Multi-Dataset Supervised Contrastive Learning (MSupCL): We read two datasets simultaneously and generated training pairs for contrastive learning. We computed the supervised contrastive loss between pairs of samples from different datasets with the same labels. This loss updates the feature extractor via gradient descent. After training, we performed classification using the updated features.

$$L^{a}_{MSupCL} = -\frac{1}{|P_a|} \sum_{v_p \in P_a} \log \frac{\exp(z_a \cdot z_p/\tau)}{\sum_{v_k \in K_a} \exp(z_a \cdot z_k/\tau)} \quad (1)$$

In the Equation 1 [3], a is an anchor from the batch. P_a means a set of all positive pairs of anchor a. K_a means a set of all positive plus negative pairs of anchor a. z_a means the feature embedding of a, z_p and z_k for one of the instance p from P_a, and one of the instance k from K_a. Finally, the τ mean the temperature.

3) Self-Supervised Contrastive Learning (SSCL): For each instance in a dataset, we applied two data augmentations to create a pair of views. We computed the NT-Xent loss between origin data and these views and updated the feature extractor via gradient descent. We then performed classification using the trained feature extractor.

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(z_i, z_k)/\tau)} \quad (2)$$

In the Equation 2 [4], Where N denotes that there is N examples from a batch, and $2N$ means that after augmentation, there will be $2N$ datapoints. Similarly, z_i, z_j, and z_k means the feature embedding of datapoint i, j,and k, the τ mean the temperature. The $\mathbb{1}_{[k \neq i]}$ is an indicator function, its value will be equal to 1 if and only if the $k \neq i$, otherwise it will be 0. The loss functions of both contrastive methods are designed to bring similar samples closer together and push different samples farther apart.

IV. Experiments

A. Datasets

1) Violence Video Dataset [8]: The Real-Life Violence Videos dataset consists of 2000 video clips from YouTube. It has two classes: Violence and NonViolence, with 1000 samples each. Violent videos include street fights and movie fights. Non-violent videos include opera performances and handcrafting.

2) TikTok Video Dataset [9]: The TikHarm dataset contains harmful videos from TikTok. It originally had four classes: Adult Content, Violence, Suicide, and Safe. We converted these into two classes: 693 samples of Harmful videos and 698 samples of Safe videos.

3) Challenges in Datasets:

- Variability: Most of the videos in the violence dataset are landscape, longer in duration, and of the same size. The TikHarm dataset, on the other hand, has a lot of vertical screen videos, which have a very diverse distribution with varying durations and varying degrees of clarity.
- Low Inter-Class Variance: Small visual differences between classes. Many videos containing harmful content differ little from safe videos, like the conflict in sports and normal sports.

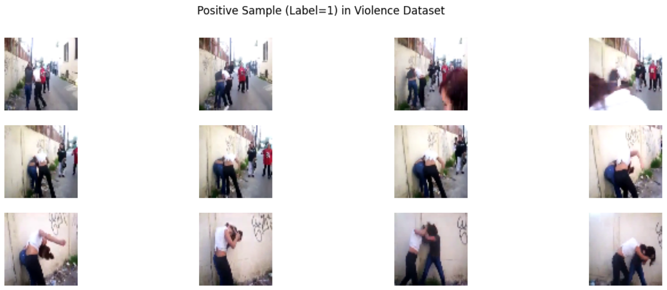


Fig. 2: Positive Sample from Violence Dataset after Resize

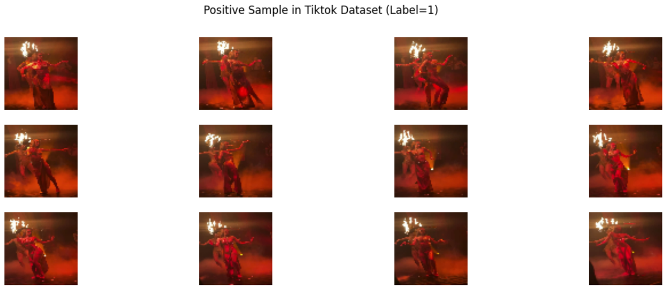


Fig. 3: Positive Sample from Tikharm Dataset after Resize

- High Intra-Class Variance: Significant differences within the same class. Safe content encompasses too wide a range of video categories that are too disparate from each other, such as making crafts and movies.

4) Data Preprocessing: We split each dataset into stratified training (55%), validation (15%), and test sets (30%). We used a batch size of 4 and an input shape of (12, 64, 64, 3), representing 12 frames of size 64×64 with 3 color channels. Keyframes were extracted by averaging frames with the same interval between each frame. If the total number of video frames is less than 12 frames, we repeat the last frame. For the MSupCL method, we formed a batch of samples from different datasets to be used to compute the contrast loss function Equation 1. Since the SSCL(SimCL) method requires data augmentation to generate training pairs, we took a number of different randomized data augmentation methods to generate data augmented views.

B. Experimental Setup

1) Hardware Specifications: This experiment was run on a personal computer with Nvidia Geforce RTX 4070 12GB, 32GB RAM. A full round of training and evaluating took up to 5 hours.

C. Results

TABLE I: Overall Accuracy Comparison for Violence and TikHarm datasets

Dataset	**C3D**			**R2+1D_18**		
	Bin-Class	**MSupCL**	**SSCL**	**Bin-Class**	**MSupCL**	**SSCL**
Violence	0.5833	0.5000	0.5000	0.5000	0.4479	0.5833
TikHarm	0.4791	0.5208	0.5208	0.5208	0.4792	0.5521

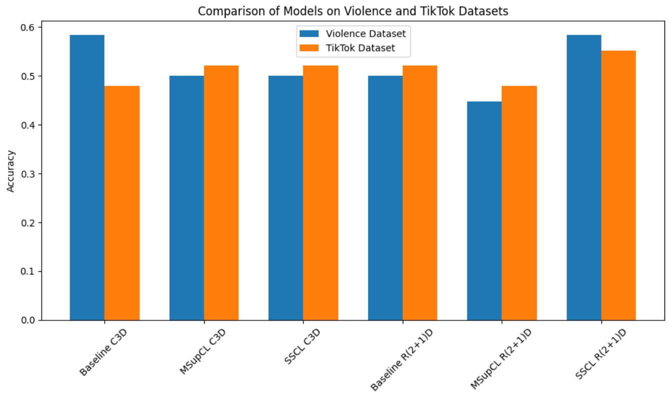
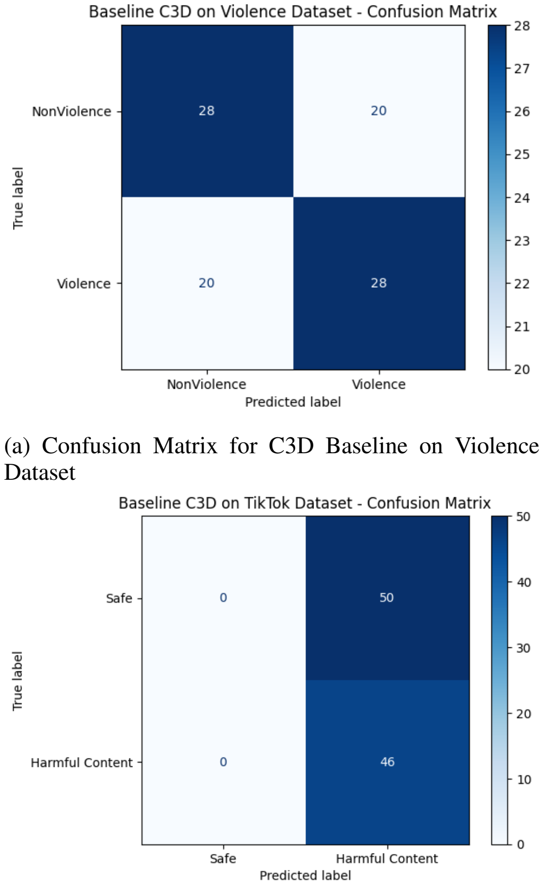


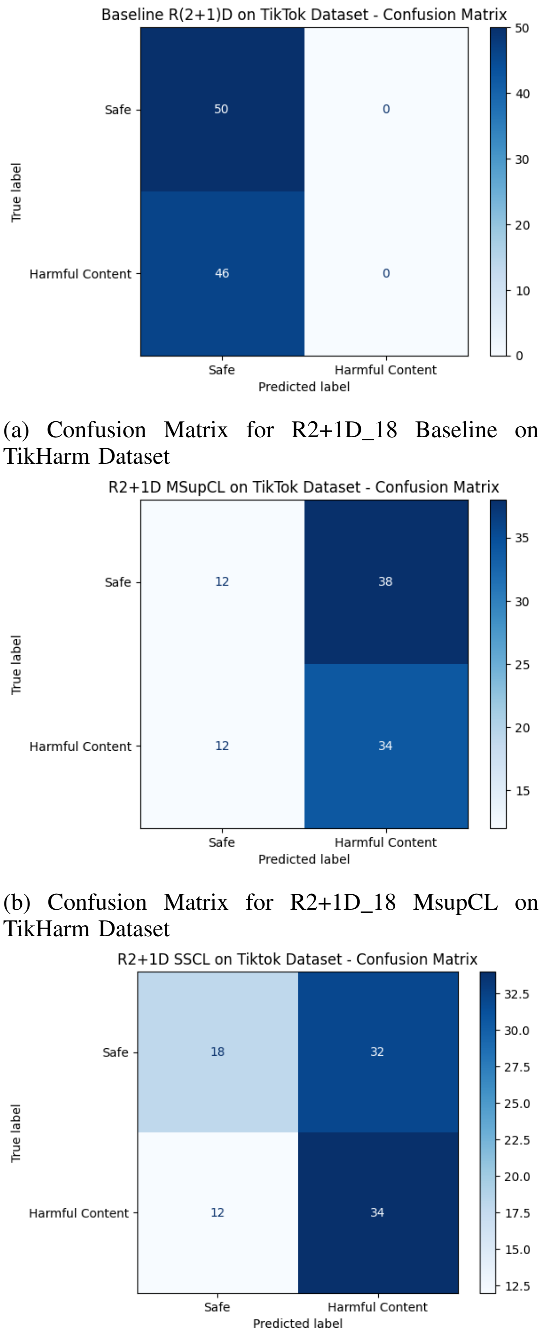
Fig. 4: Accuracy Comparison in two datasets



(a) Confusion Matrix for C3D Baseline on Violence Dataset

(b) Confusion Matrix for C3D Baseline on TikHarm Dataset

Fig. 5: Comparison of Performance of Baseline between Violence and TikHarm Datasets



(a) Confusion Matrix for R2+1D_18 Baseline on TikHarm Dataset

(b) Confusion Matrix for R2+1D_18 MsupCL on TikHarm Dataset

(c) Confusion Matrix for R2+1D_18 SSCL on TikHarm Dataset

Fig. 6: Comparison of Performance of Three methods on the TikHarm Datasets

D. Analysis

The Baseline C3D model achieved around 58% accuracy on the violence dataset and 48% on the TikTok dataset. The MSupCL and SSCL methods with the C3D extractor did not significantly improve accuracy. As shown in Figure 5, the TikHarm dataset is more difficult to classify in the Baseline method using C3D as the feature extractor. All the samples were classified as positive.

For the R(2+1)D extractor, the SSCL method achieved about 58% accuracy on the violence dataset and 55% on the TikTok dataset, showing slight improvement. And Figure 6 demonstrates that the baseline approach classified all models in the TikHarm dataset as negative. Whereas both MSupCL and SSCL succeeded in making the prediction results more diverse, we believe that they succeeded in allowing the models to extract characteristics of the two types of data differently.

The difficulty in distinguishing categories for the baseline model can be seen through these confusion matrices. The model sometimes only predicted one category, which suggests that the model might not be able to distinguish features effectively.

E. Discussion

The MsupCL Contrastive Learning approach did not significantly improve classification performance in this setting. Factors causing this may include

- Datasets chosen: The representations of the two datasets we chose may not help much in categorizing each other.
- Challenging dataset: Low inter-class variance and high intra-class variance.
- Limited data: Small size of the sampled dataset, which may not satisfy the requirements of the deep learning model.
- Hyperparameters: Epoch, learning rate, batch size and temperature may need to be tuned for better performance.

V. Conclusion

We explored the impact of contrastive learning on harmful content video classification. We compared a baseline method, a multi-dataset supervised contrastive learning approach, and a self-supervised contrastive learning approach using two feature extractors.

Our experiments showed that MSupCL contrastive learning did not significantly improve performance in this setting. The challenges posed by the datasets, may have affected the feature extractors' ability to learn effective representations.

Further research is needed to understand how to apply contrastive learning effectively in this domain.

VI. Future Work

Future research could focus on:

- Larger datasets: using more training data to train the feature extractor
- Sharper video frames: using video frames with more pixels to allow the network to extract features
- Improved comparison methods: using larger batch for comparison training.
- Pre-trained models: Fine-tune models trained on large action recognition datasets.

References

[1] A. Arora, P. Nakov, M. Hardalov, S. M. Sarwar, V. Nayak, Y. Dinkov, D. Zlatkova, K. Dent, A. Bhatawdekar, G. Bouchard, and I. Augenstein, "Detecting harmful content on online platforms: What platforms need vs. where research efforts go," *ACM Comput. Surv.*, vol. 56, Oct. 2023.

[2] N. Mumtaz, N. Ejaz, S. Habib, S. M. Mohsin, P. Tiwari, S. S. Band, and N. Kumar, "An overview of violence detection techniques: current challenges and future directions," *Artificial Intelligence Review*, vol. 56, p. 4641–4666, Oct. 2022.

[3] A. Rani and Y. Verma, "Activity-based early autism diagnosis using a multi-dataset supervised contrastive learning approach," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7788–7797, January 2024.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. Singh, eds.), vol. 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, PMLR, 13–18 Jul 2020.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, (USA), p. 4489–4497, IEEE Computer Society, 2015.

[6] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

[7] F. U. M. Ullah, M. S. Obaidat, A. Ullah, K. Muhammad, M. Hijji, and S. W. Baik, "A comprehensive review on vision-based violence detection in surveillance videos," *ACM Comput. Surv.*, vol. 55, Feb. 2023.

[8] M. M. Soliman, M. H. Kamal, M. A. El-Massih Nashed, Y. M. Mostafa, B. S. Chawky, and D. Khattab, "Violence recognition from videos using deep learning techniques," in *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pp. 80–85, 2019.

[9] A. H. Vo, "Tikharm dataset," Jun 2024.