QEBB: A Query-Efficient Black-Box Adversarial Attack on Video Recognition Models Based on Unsupervised Key Frame Selection

Kimia Haghjooei[a] and Mansoor Rezghi[b]
Department of Computer Science, Tarbiat Modares University, Tehran, Iran

Keywords: Adversarial Examples, Adversarial Attack, Video Recognition, Black-Box Attack.

Abstract: Despite the success of deep learning models, they remain vulnerable to adversarial attacks introducing slight perturbations to inputs, resulting in adversarial examples. Black-box attacks, where model details are hidden from the attacker, gain attention for their real-world applications. Although studying adversarial attacks on video models is crucial due to their surveillance importance and security applications, most works on adversarial examples mainly focus on images, and videos are rarely studied since attacking videos is more challenging. Recent black-box video attacks involve selecting key frames to reduce video's dimensionality. This addresses the high costs of attacking the entire video but may require numerous queries, making the attack noticeable. Our work introduces QEBB, a query-efficient black-box video attack. We employ an unsupervised key frame selection method to choose frames with vital representative information. Using saliency maps, we focus on key frame salient regions. QEBB successfully attacks UCF-101 and HMDB-51 datasets with 100% success and reducing query numbers by nearly 90% in comparison to state-of-the-art methods.

1 INTRODUCTION

Deep Neural Networks have demonstrated a lot of power and success across various computer vision tasks such as image classification (Jiang et al., 2023; Mittal et al., 2022; Paymode and Malode, 2022), video recognition (Surek et al., 2023; Pham et al., 2022; Wu et al., 2023), face recognition (Boussaad and Boucetta, 2022; Kurakin et al., 2018; Li et al., 2022) and object detection (Ajagbe et al., 2022; Zaidi et al., 2022). Despite their success, deep neural networks have shown vulnerability to adversarial examples (Goodfellow et al., 2014). Recent studies have shown that adding a slight perturbation to a clean input can fool a DNN and lead to incorrect output. Therefore, studying adversarial attacks is critical, especially since deep learning models are used in security-critical applications (Nasir et al., 2022; Arunnehru et al., 2023).

Adversarial examples can be generated by an adversarial attack in either a white-box manner (Wang et al., 2022; Agnihotri and Keuper, 2023; Carlini and Wagner, 2017; Moosavi-Dezfooli et al., 2016), where the attacker has full knowledge of the model's structure or a in a black-box manner (Cheng et al., 2018; Carlini and Wagner, 2018; Zhang et al., 2022; Wan et al., 2023), where the attacker has no information about the model and can only query predictions for specific inputs. Therefore, black-box attacks seem to make more realistic assumptions. However, it is worth mentioning that a high number of queries would make the attack visible to defence mechanisms.

Furthermore, adversarial attacks can be categorized as targeted attacks (Sadrizadeh et al., 2023), where the goal is to make the model predict a specific adversarial label or untargeted attacks (Zhou et al., 2022), where the output of model after attack is not important as long as it differs from the original label.

Most of the recent research on adversarial examples have been considered on image models and adversarial attacks on video recognition models have been rarely studied. Since video classification models are used in surveillance applications (Sultani et al., 2018), it is necessary to assess their robustness towards adversarial attacks. Although several white-box (Wei et al., 2019; Li et al., 2018; Pony et al., 2021; Lo and Patel, 2021) adversarial attacks have been proposed for video recognition models, black-box attacks on these models have been rarely studied. PATCHATTACK (V-BAD) (Jiang et al., 2019) represents the first attempt to design a black-box attack on

[a] https://orcid.org/0009-0001-1722-9739
[b] https://orcid.org/0000-0003-4214-5008

Haghjooei, K. and Rezghi, M.
QEBB: A Query-Efficient Black-Box Adversarial Attack on Video Recognition Models Based on Unsupervised Key Frame Selection.
DOI: 10.5220/0012359900003654

In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2024)*, pages 288-295
ISBN: 978-989-758-684-2; ISSN: 2184-4313

video recognition models. It initially uses a local image classifier to generate perturbations for each video frame, and then updates the perturbations by querying the target model. To design adversarial attacks for video recognition models, one approach involves treating videos as sets of images and apply existing attacks proposed for image models. Nevertheless, this method can be significantly time-consuming due to the higher dimensionality of videos compared to images. To address this issue, a potential solution involves reevaluating the approach for designing black-box adversarial attacks on video recognition models. A novel aspect of this approach is to leverage the spatial and temporal redundancies present in video data to reduce the complexity of generating adversarial video examples. One potential method is to reduce the dimensionality of video data by selecting a subset of frames as key frames and performing the adversarial attack on this specific set of frames rather than on all frames. Key frames are defined as frames that contribute the most to representing the actual video and play a crucial role in the classification task. Recent black-box adversarial attacks on video recognition models have introduced innovative attacks that include key frame selection before the attack phase. For example, the Heuristic attack by Zhang et al. (Wei et al., 2020) estimates the importance of each frame based on its role in the classification task, although it requires a high number of queries. Furthermore, they utilize a saliency map to target the important regions of key frames during the attack process. Despite its effectiveness, this method is still time-consuming and demands a significant number of queries, which is an important criterion when evaluating black-box adversarial attacks, as a high number of queries can make the attack more visible. Additionally, Wei et al. introduced the SVA attack (Wei et al., 2022), which employs reinforcement learning to select key frames. Despite its novelty, this attack also requires a substantial number of queries during the key frame selection phase.

Therefore, recent advancements in black-box adversarial examples have effectively addressed the challenge posed by high-dimensional video data through the introduction of a key frame selection process (Wei et al., 2020; Wei et al., 2022). However, it's important to note that their key frame selection methods (Wei et al., 2020; Wei et al., 2022) heavily rely on the classifier model, determining a frame's importance based on its impact on the classification outcome. This approach presents two major issues: Firstly, these selection processes necessitate a substantial number of queries to compute individual scores for each frame, indicating its influence on the classification result. Secondly, these methods assess the importance of each frame independently and select those with the highest classification scores as key frames. Consequently, these methods do not explore potential sets of frames, potentially missing out on possible candidates for key frames that could be crucial.

To tackle this issue, we propose a novel approach for designing a black-box adversarial attack on video recognition models, based on the Heuristic attack (Wei et al., 2020). In our method, we introduce an unsupervised key frame selection process and redefine what constitutes key frames. We define key frames as the set of frames containing the most informative details about a video. These frames contribute significantly to the classification process, as they signify essential representative characteristics of the video. Moreover, these key frames represent the entire video data, encapsulating its overall information. To select such frames in an unsupervised manner, we employ k-means, a simple clustering technique. Our work results in a significantly reduced number of queries while achieving a 100% fooling rate on two benchmark datasets. In summary, our main contributions are as follows:

- We study the problem of black-box adversarial attacks on video recognition models and proposed an untargeted query-efficient black-box attack called QEBB.
- For each clean video, we define a subset of frames as key frames using an unsupervised selection method. We generate video adversarial examples by perturbing key frames only.
- We conducted a series of comprehensive experiments on two benchmark video datasets and a video recognition model indicating that our method not only requires a significantly smaller number of queries, but also generates more realistic adversarial examples closely resembling real videos.

2 RELATED WORK

In this section we review popular adversarial attacks on both image and video models.

2.1 Adversarial Attack on Image Models

Recent studies on adversarial examples are mainly focused on image classification models. Various attacks

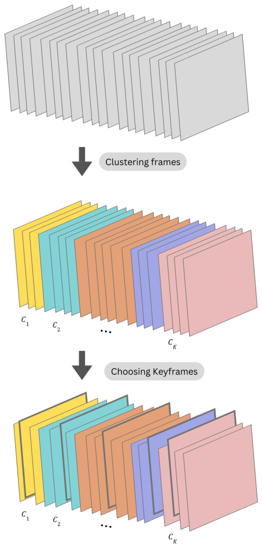

Figure 1: Selecting key frames using a clustering algorithm. After collecting frames into K clusters, the nearest frames to the cluster center are chosen as key frame candidates.

are designed in both white-box attacks (Wang et al., 2022; Agnihotri and Keuper, 2023; Carlini and Wagner, 2017; Moosavi-Dezfooli et al., 2016) and black-box (Cheng et al., 2018; Carlini and Wagner, 2018; Zhang et al., 2022; Wan et al., 2023) manners. In the Opt-attack (Cheng et al., 2018), θ represents the search direction, and the distance from a clean image x to the decision boundry along θ is defined by $g(\theta)$. The goal in Opt-attack (Cheng et al., 2018)is to minimize $g(\theta)$.

2.2 Adversarial Attack on Video Models

The number of existing adversarial attacks on video models is significantly lower than the efforts made for image models duo for several reasons. Firstly, videos consist of high-dimensional data making them more complex. Secondly, applying existing adversarial attacks designed for image models to video models consumes a lot of time, resources and a large amount of queries, making the attack more detectable from a security standpoint. Hence, it is crucial to design attacks for video models specifically.

Recent adversarial attacks on video recognition models are mainly in a white-box manner. For instance, (Wei et al., 2019) discusses the sparsity of adversarial perturbations through frames. Li et al. (Li et al., 2018) proposed a novel approach to produce perturbation clips in order to achieve higher attack success rate. Pony et al. (Pony et al., 2021) proposed Flickering attack that is generalized to make universal perturbations. Lo et al. (Lo and Patel, 2021) proposed MultAV which generates perturbations on videos by using multiplication.

Whilst several white-box adversarial attacks have been proposed on video recognition models, attacks in a black-box setting have been rarely studied. Jiang et al. (Jiang et al., 2019) claimed to be the first attempt to design a black-box attack on video models called V-BAD which generates initial perturbations for each video frame utilizing a local image classifier, and then updates the perturbations by querying the target model. Compared to V-BAD, our work doesn't require a local image classifier and requires significantly lower number of queries. One of the state-of-art methods in this field is the Heuristic attack (Wei et al., 2020), which introduces an innovative approach to tackle the challenges associated with high-dimensional video data. This method involves selecting a small subset of key frames for each input video. Specifically, Heuristic ranks frames based on their classification scores and chooses those with the highest scores. Subsequently, salient regions within these selected key frames are targeted with perturbations. In fact, Heuristic attack (Wei et al., 2020) defines frames as key frames based on their impact on the discrimination task. Despite its success in deceiving video models, Heuristic attack (Wei et al., 2020) still demands a substantial number of queries to identify suitable key frames, potentially making the attack more noticeable in a black-box setting. Another novel attack, SVA (Wei et al., 2022), employs reinforcement learning to select key frames. While effective in attacking video classification models, this method also requires a significant number of queries for its frame selection process.

In this work, we approach the key frame selection process differently. Specifically, we look for frames that can best represent the video and its features in an unsupervised manner. We choose the most representative and important frames, which make the most significant contribution to the classification process due to their rich informational content about the video. Our approach achieves far fewer queries compared to state-of-the-art black-box attacks.

3 PROPOSED METHOD

We indicate a video recognition model as a function f. Specially, $f(x)$ takes a clean video $X \in \mathbb{R}^{T\times W\times H\times C}$ as an input and outputs $\hat{y}$ as its top-1 class and the corresponding probability $P(\hat{y}|X)$ where T,W,H,C denote the number of frames, width, height and the number of channels respectively. The true class $y \in Y = \{1, 2, \cdots, V\}$ where V is the number of classes. The adversarial example X_{adv} is resulted by perturbing the original video X. In the untargeted setting we aim to make $f(X_{adv}) \neq y$. Recent black-box attacks (Wei et al., 2020; Wei et al., 2022) have introduced various key frame selection methods to generate more efficient video adversarial examples by targeting key frames rather than all frames of a video. Although they managed to produce adversarial examples with higher qualities, they require a substantial number of queries to select key frames. Therefore, we propose an unsupervised key frame selection approach which enhances attack efficiency bu eliminating the need for a high number of queries, making it more practical for real-world scenarios. In fact, we select a subset of key frames and conduct the QEBB attack, inspired by Opt-attack (Cheng et al., 2018), on these specific frames rather than on all frames. Specifically, if we indicate the selected key frames as $\hat{X}$ and other frames as $X^{'}$, we can represent video $X = (\hat{X}, X^{'})$. Therefore, we define the following function to indicate the output of the model:

$$h(\hat{X}) = f(X) = y \tag{1}$$

There are multiple methods to define and select key frames. In this paper, we propose a novel key frame selection method from a different point of view. Specifically, if we consider a video X as a collection of frames followed by a time dimension which are temporally related, frames that are temporally close to each other share a high degree of similarity. In fact, we can categorize all of the frames based on their similarities into some groups so that frames within a group would be highly similar, and frames from different groups would be different. Therefore, by choosing some frames from each group as the indicator of the frames of that group, we would be able to find a key frame set that has the most representative details of the video while demonstrating the overall flow of the video. For this manner, we have used k-means, a simple clustering algorithm. Specifically, consider $\Psi = \{x_i = X(i,:,:,:) | i = 1,2,...,T\}$ $(\Psi = \hat{X} \bigcup X^{'})$ as the frame set of video X where x_i demonstrates the ith frame of video X. Consequently, after performing k-means on Ψ, we obtain K clusters $C = \{C_1, C_2, ..., C_K\}$. We select a subset of frames for each cluster that are closer to the center of the cluster and represent the members of that cluster in the best possible way. Therefore, such a statement can be formulated as follows:

$$\chi_j = \underset{x \in C_j}{\arg\min} \|x - \mu_j\|_F^2 \tag{2}$$

where χ_j is the set of representative frames for C_j, indicating the jth cluster, and μ_j denotes the center of it. Hence, we choose a set of frames for each cluster as the representative of the members of that cluster. Therefore, we construct the key frames set as follows:

$$\hat{X} = \bigcup_{j=1}^{k} \{\chi_j\} \tag{3}$$

Figure1 indicates a schema of our key frame selection.

Opt-attack (Cheng et al., 2018) defines a direction θ and searches for the closest distance $g(\theta)$ where an adversarial example can be found. For further efficiency, Opt-attack (Cheng et al., 2018) improves θ iteratively. In this paper, we extend the Opt-attack (Cheng et al., 2018) to video recognition models. Specially, same as Opt-attack (Cheng et al., 2018), our objective is to find

$$\min_{\theta} g(\theta) \tag{4}$$

where $g(\theta)$ is defined as

$$g(\theta) = \min_{\lambda} \left(h(\hat{X} + \frac{\theta}{||\theta||}.\lambda) \neq y \right) \tag{5}$$

meaning that we perform attack on key frames only.

Therefore, we conduct our attack on $\hat{X}$. Moreover, as in (Wei et al., 2020), we used a saliency map (Lee et al., 2012) for further efficiency. Hence, if we consider the salient region mask as M, X_{adv} is generated as below:

$$X_{adv} = \left(\left(\hat{X} + \theta^{\star}.g(\theta)^{\star} \right) * M, X^{'} \right) \tag{6}$$

Moreover, same as Heuristic attack (Wei et al., 2020), we initialize direction $\theta = \frac{p}{||p||}$ where $p = Z - X$ and Z is an input video which comes from a different class.

Finally, we obtain the adversarial example $X_{adv} = X + g(\theta^{\star}) \times \theta^{\star}$ where $\theta^{\star}$ is the optimal solution in an iterative manner by updating θ. For this aim, we use Zero-Order-Optimizatin method (Chen et al., 2017), a method that defines the estimated gradient as follows:

$$g^{'} = \frac{g(\theta + \beta u) - g(\theta)}{\beta}.u \tag{7}$$

Here, u represents a random Gaussian vector with same dimensions as θ. Moreover, $\beta > 0$ indicates a smoothing parameter which is subjected to a tenfold reduction if the estimated gradients fail to offer meaningful insights for the updating of θ (Wei et al., 2020). Hence, we update θ in each iteration as follows:

$$\theta \leftarrow \theta - \alpha.g^{'} \tag{8}$$

where α indicates the step size of each iteration. In conclusion, we tackle the challenges of generating efficient video adversarial examples using an unsupervised manner which results in a significant reduction of number of queries required for this process while selecting the best potential candidates for key frames containing the overall informative details of a video.

Table 1: Numerical evaluation of our framework, QEBB(including its two variations *N*-QEBB and α-QEBB), compared to state-of-art black-box attacks.

Model	Dataset	Attack	MQ	MT	MAP	MSSIM	FR(%)
C3D	HMDB51	*N*-QEBB	**349.21**	10.255	95.776	0.097	**100**
		α-QEBB	**349.82**	9.225	90.213	**0.1036**	**100**
		Heuristic(Wei et al., 2020)	5947.9	11.05	94.32	0.101	**100**
		SVA(Wei et al., 2022)	3328.9	**4.67**	**56.84**	17e-5	**100**
		VBAD(Jiang et al., 2019)	68584.2	32.24	59.59	76e-6	95
	UCF101	*N*-QEBB	**352.21**	10.657	96.22	**0.0423**	**100**
		α-QEBB	**357.79**	11.45	96.23	**0.0381**	**100**
		Heuristic(Wei et al., 2020)	53596.4	55.54	96.15	0.022	**100**
		SVA(Wei et al., 2022)	4473.8	**7.11**	**53.24**	12e-5	89
		VBAD(Jiang et al., 2019)	71480.8	31.68	56.5	52e-6	87

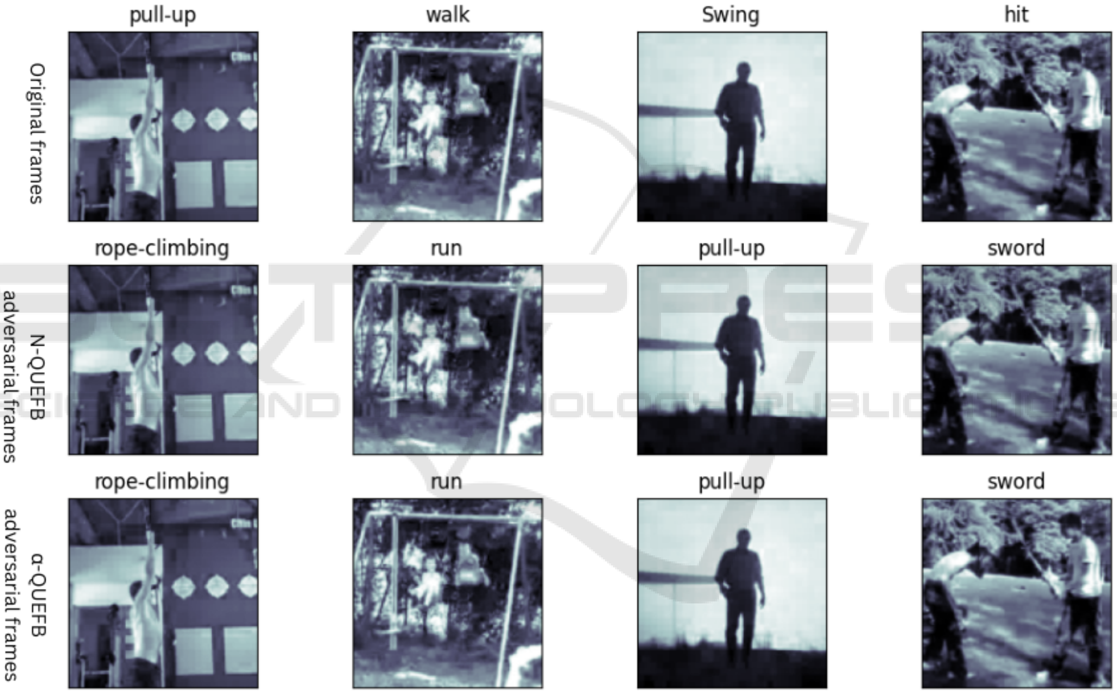

Figure 2: Examples of adversarial frames generated by *N*-QEBB and α-QEBB under the untargeted setting.

4 EXPERIMENTS

In this section, we provide a comprehensive evaluation to test the performance of our proposed quarry-efficient, untargeted, black-box adversarial attack on two benchmark video datasets. Our evaluation contains various aspects, including the reduction in overall perturbation, significant decrease in query numbers required for the attack, resulting in adversarial examples that are highly imperceptible to the human eye. Furthermore, we offer a detailed assessment of our method, showcasing its efficiency and effectiveness among state-of-art black-box attacks on video recognition models.

4.1 Experiment Setting

Datasets. We used two common video datasets for our evaluation: UCF-101 (Soomro et al., 2012) and HMDB-51 (Kuehne et al., 2011). UCF-101 is an action recognition datasets derived from YouTube that contains 13,320 videos with 101 action classes.

HMDB-51 is a large human motion dataset which contains 7000 videos with 51 action categories. For both datasets, we used 70% of videos for training set and the rest of 30% for test set as in (Wei et al., 2020). In (Wei et al., 2020) for each video, 16-frame snippets are extracted using uniform sampling. In our setting, we then perform our novel key frame selection method on these 16 frames.

Video Recognition Models. We used C3D (Hara et al., 2018), a popular video recognition model as our target models. Moreover, we consider that the attacker can only access top-1 class and its corresponding probability.

Parameter Setting. As in (Wei et al., 2020), the parameter tuning is done on 30 videos that are randomly sampled from the test set. We set the area ratio of salient region ϕ to 0.6. We also set the step size for updating the gradient to 0.2 on UCF-101. On the other hand, we set a larger step size for some samples on HMDB-51. Moreover, we set the number of cluster to $K = 5$. Furthermore, after clustering frames, we choose the representative frames for each cluster as either N-nearest frame (N-QEBB) or to the α% nearest frames (α-QEBB) to the cluster center. In our evaluation, we set $\alpha = 20\%$ and $N = 1$.

4.2 Evaluation Metrics

We employ five metrics to comprehensively assess our method's performance:

Fooling Rate (FR). This metric represents the ratio of successfully misclassified adversarial videos.

MT (Average Running Time). It denotes the average time in minutes required to execute the attack on test samples.

MQ (Average Query Number). This metric signifies the average number of queries necessary to generate each adversarial example.

MAP (Mean Absolute Perturbation). This metric indicates the mean perturbation in each pixel throughout the entire video:

$$MAP = \frac{1}{N}\sum_{i}^{N}\frac{||x_{i,adv} - x_i||}{|P_i|} \tag{9}$$

where N denotes the number of test samples and P_i represents the total number of pixels existing in x_i.

MSSIM (Mean Structural Similarity Index Measure). It quantifies the average SSIM similarity between each adversarial example and its corresponding clean video:

$$SSIM(x^j, x^j_{adv}) = \frac{(2\mu_{x^j_{adv}}\mu_{x^j_{adv}} + C_1)(2\sigma_{x^j, x^j_{adv}} + C_2)}{(\mu^2_{x^j} + \mu^2_{x^j_{adv}} + C_1)(\sigma^2_{x^j} + \sigma^2_{x^j_{adv}} + C_2)} \tag{10}$$

where x^j indicates the jth frame of the video x while μ_{x^j} and $\sigma^2_{x^j}$ are showing the mean and the variance of the jth frame of x respectively. Furthermore, $\sigma_{x^j, x^j_{adv}}$ shows the correlation coefficient between the jth frame of x_{adv} and x. Moreover, C_1 and C_2 are numeric parameters.

These metrics collectively provide a comprehensive evaluation of our method's effectiveness and efficiency.

4.3 Performance Evaluation

In our comprehensive evaluation, we achieved a comparative analysis between our proposed methods, namely N-QEBB and α-QEBB, and three state-of-the-art black-box attacks: Heuristic (Wei et al., 2020), SVA (Wei et al., 2022), and VBAD (Jiang et al., 2019). Our assessments were carried out using a single video recognition model on two benchmark datasets, UCF-101 (Soomro et al., 2012) and HMDB-51 (Kuehne et al., 2011). The results, presented in Table 1 and visually illustrated in Figure 2, provide compelling evidence of the effectiveness and efficiency of our framework.

Table 1 indicates the superior performance of both variations of our method. N-QEBB and α-QEBB both achieve a 100% fooling rate on both datasets, outperforming other approaches. particularly, they achieve this with significantly fewer queries, a crucial factor when evaluating the stealthiness of black-box attacks from a security perspective.

Moreover, while we achieve a significant reduction in the number of queries on both datasets, both N-QEBB and α-QEBB succeed in generating adversarial examples with higher Mean Structural Similarity Index Measure (MSSIM). This indicates that our adversarial examples closely resemble the clean videos compared to other attacks.

In terms of computational time, both of our frameworks require less time compared to Heuristic (Wei et al., 2020) and VBAD (Jiang et al., 2019). Although N-QEBB and α-QEBB are more time-consuming compared to SVA (Wei et al., 2022) attack, they achieve fully successful attacks on both datasets.

Furthermore, it is worth mentioning that α-QEBB generates adversarial examples with lower Mean Absolute Perturbation (MAP) compared to Heuristic (Wei et al., 2020) attack on HMDB. However, MAP is not necessarily the most suitable metric for assessing video adversarial examples since it does not account for the spatial and temporal relations within video data. In contrast, other criteria such as MSSIM are more suitable for comparing video examples since they consider the spatial relations of video frames.

Both of our frameworks on UCF and α-QEBB on HMDB achieve higher MSSIM compared to the other attacks, indicating that our frameworks generate more similar adversarial examples to clean videos.

Moreover, even though SVA (Wei et al., 2022) stands out as one of the most effective existing attacks with faster execution and superior Mean Absolute Perturbation (MAP) scores, our framework outperforms SVA (Wei et al., 2022) by requiring significantly lower number of queries, along with achieving 100% successful adversarial examples and higher Mean Structural Similarity Index Measure (MSSIM).

Figure 2 visually indicates examples of adversarial examples generated by both *N*-QUEFB and α-QUEFB. As shown in this figure, our frameworks can produce adversarial examples that closely resemble the original videos, supporting the quality of our approach.

In conclusion, our proposed unsupervised key frame selection method combined with saliency-based perturbations, significantly enhances attack efficiency. This method reduces the need for queries, making our framework less detectable from a security standpoint, while generating high-quality adversarial examples. These findings emphasize the potential of our approach in the black-box adversarial attacks on video recognition models.

5 CONCLUSIONS

Generating video adversarial examples poses a significant challenge due to their high-dimensional nature. To enhance efficiency, recent black-box attacks target only a subset of the video frames as keyframes. This approach, while more efficient, often requires many queries, making attacks detectable by defense mechanisms, as keyframes are chosen based on their impact on recognition tasks. In this paper, we introduce an innovative, unsupervised keyframe selection method using simple clustering, where we group frames and select representative frames from each group as keyframes. This method significantly reduces the number of required queries, enabling the generation of imperceptible adversarial examples by focusing on representativeness rather than influence on recognition tasks. Future work may explore various video summarization techniques for more effective keyframe selection.

REFERENCES

Agnihotri, S. and Keuper, M. (2023). Cospgd: a unified white-box adversarial attack for pixel-wise prediction tasks. *arXiv preprint arXiv:2302.02213*.

Ajagbe, S. A., Oki, O. A., Oladipupo, M. A., and Nwanakwaugwu, A. (2022). Investigating the efficiency of deep learning models in bioinspired object detection. In *2022 International conference on electrical, computer and energy technologies (ICECET)*, pages 1–6. IEEE.

Arunnehru, J. et al. (2023). Deep learning-based real-world object detection and improved anomaly detection for surveillance videos. *Materials Today: Proceedings*, 80:2911–2916.

Boussaad, L. and Boucetta, A. (2022). Deep-learning based descriptors in application to aging problem in face recognition. *Journal of King Saud University-Computer and Information Sciences*, 34(6):2975–2981.

Carlini, N. and Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee.

Carlini, N. and Wagner, D. (2018). Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*, pages 1–7. IEEE.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. (2017). Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26.

Cheng, M., Le, T., Chen, P.-Y., Yi, J., Zhang, H., and Hsieh, C.-J. (2018). Query-efficient hard-label black-box attack: An optimization-based approach. *arXiv preprint arXiv:1807.04457*.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Hara, K., Kataoka, H., and Satoh, Y. (2018). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555.

Jiang, H., Diao, Z., Shi, T., Zhou, Y., Wang, F., Hu, W., Zhu, X., Luo, S., Tong, G., and Yao, Y.-D. (2023). A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation. *Computers in Biology and Medicine*, page 106726.

Jiang, L., Ma, X., Chen, S., Bailey, J., and Jiang, Y.-G. (2019). Black-box adversarial attacks on video recognition models. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 864–872.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE.

Kurakin, A., Goodfellow, I. J., and Bengio, S. ((2018)). Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC.

Lee, Y. J., Ghosh, J., and Grauman, K. (2012). Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1346–1353. IEEE.

Li, M., Huang, B., and Tian, G. (2022). A comprehensive survey on 3d face recognition methods. *Engineering Applications of Artificial Intelligence*, 110:104669.

Li, S., Neupane, A., Paul, S., Song, C., Krishnamurthy, S. V., Chowdhury, A. K. R., and Swami, A. (2018). Adversarial perturbations against real-time video classification systems. *arXiv preprint arXiv:1807.00458*.

Lo, S.-Y. and Patel, V. M. (2021). Multav: Multiplicative adversarial videos. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE.

Mittal, S., Srivastava, S., and Jayanth, J. P. (2022). A survey of deep learning techniques for underwater image classification. *IEEE Transactions on Neural Networks and Learning Systems*.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582.

Nasir, I. M., Raza, M., Shah, J. H., Wang, S.-H., Tariq, U., and Khan, M. A. (2022). Harednet: A deep learning based architecture for autonomous video surveillance by recognizing human actions. *Computers and Electrical Engineering*, 99:107805.

Paymode, A. S. and Malode, V. B. (2022). Transfer learning for multi-crop leaf disease image classification using convolutional neural network vgg. *Artificial Intelligence in Agriculture*, 6:23–33.

Pham, H. H., Khoudour, L., Crouzil, A., Zegers, P., and Velastin, S. A. (2022). Video-based human action recognition using deep learning: a review. *arXiv preprint arXiv:2208.03775*.

Pony, R., Naeh, I., and Mannor, S. (2021). Over-the-air adversarial flickering attacks against video recognition networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 515–524.

Sadrizadeh, S., Aghdam, A. D., Dolamic, L., and Frossard, P. (2023). Targeted adversarial attacks against neural machine translation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Soomro, K., Zamir, A. R., and Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Sultani, W., Chen, C., and Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488.

Surek, G. A. S., Seman, L. O., Stefenon, S. F., Mariani, V. C., and Coelho, L. d. S. (2023). Video-based human activity recognition using deep learning approaches. *Sensors*, 23(14):6384.

Wan, J., Fu, J., Wang, L., and Yang, Z. (2023). Bounceattack: A query-efficient decision-based adversarial attack by bouncing into the wild. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 68–68. IEEE Computer Society.

Wang, Y., Liu, J., Chang, X., Rodríguez, R. J., and Wang, J. (2022). Di-aa: An interpretable white-box attack for fooling deep neural networks. *Information Sciences*, 610:14–32.

Wei, X., Yan, H., and Li, B. (2022). Sparse black-box video attack with reinforcement learning. *International Journal of Computer Vision*, 130(6):1459–1473.

Wei, X., Zhu, J., Yuan, S., and Su, H. (2019). Sparse adversarial perturbations for videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8973–8980.

Wei, Z., Chen, J., Wei, X., Jiang, L., Chua, T.-S., Zhou, F., and Jiang, Y.-G. (2020). Heuristic black-box adversarial attacks on video recognition models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12338–12345.

Wu, W., Sun, Z., and Ouyang, W. (2023). Revisiting classifier: Transferring vision-language models for video recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2847–2855.

Zaidi, S. S. A., Ansari, M. S., Aslam, A., Kanwal, N., Asghar, M., and Lee, B. (2022). A survey of modern deep learning based object detection models. *Digital Signal Processing*, 126:103514.

Zhang, J., Li, B., Xu, J., Wu, S., Ding, S., Zhang, L., and Wu, C. (2022). Towards efficient data free black-box adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15115–15125.

Zhou, C., Wang, Y.-G., and Zhu, G. (2022). Object-attentional untargeted adversarial attack. *arXiv preprint arXiv:2210.08472*.