# BEYOND THE PIXELS: EXPLORING THE EFFECTS OF BIT-LEVEL VIDEO CORRUPTIONS ON MODEL PERFORMANCE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We investigate the robustness of video models under bit-level file and network corruptions, which can arise naturally from hardware errors during storage or network failures in transmission. Previous work in video robustness has focused on pixel-space corruptions, but bit-level corruptions remain relatively underexplored. Specifically, we study the effect of bit-level corruptions on video model performance, and explore strategies to defend against file and network corruptions. Our experiments demonstrate that bit-level corruptions cause substantial performance drops on the action recognition and multi-object tracking tasks. Motivated by this, we explore *corruption-agnostic* and *corruption-aware* defenses against bit-level corruptions on video ML models. The best corruption-agnostic defense, out-of-distribution detection, reaches up to 98.5% AUROC for discriminating corrupted videos at high corruption levels, while our method, Bit-corruption Augmented Training (BAT), a corruption-aware technique, provides up to 7.1% accuracy improvement compared to a no-defense baseline.

## 1 INTRODUCTION

Video is becoming an increasingly common data modality, with applications in online conferencing (Jansen et al., 2018), autonomous vehicles (Chen et al., 2018; Bojarski et al., 2016), action recognition (Kuehne et al., 2011; Soomro et al., 2012; Carreira & Zisserman, 2017), and event detection (Fu et al., 2019; Gaidon et al., 2013). Recent work in computer vision has studied the robustness of machine learning models to pixel-space corruptions such as adversarial examples (Madry et al., 2017; Carlini & Wagner, 2016; Goodfellow et al., 2014; Moosavi-Dezfooli et al., 2016; Szegedy et al., 2013). However, in the real world, videos are susceptible to corruptions beyond the pixels, such as bit-level *file and network corruptions* (see Figure 1).

Bit-level corruptions are generally non-adversarial, arising from hardware errors during storage (Sivathanu et al., 2005; Zhang et al., 2010) or network failures in transmission (Mushtaq & Mellouk, 2017). These corruptions spuriously modify bits in a file, leading to data loss or distortion (*e.g.*, objects duplicating, noisy patches, or freeze-frames). Previous bitstream-aware works in computer vision include studies of JPEG compression (Aydemir et al., 2018; Dziugaite et al., 2016; Das et al., 2018) or bit-depth modification (Guo et al., 2017b) as adversarial defenses. However, the impact of erroneous data corruptions at the bit-level on video model performance is not well understood. Moreover, it is unclear how to improve the robustness of video models against bit-level corruptions.

In this work, we take a first step in understanding and improving video model robustness to file and network corruptions. We evaluate model performance under these corruptions on action recognition (Kuehne et al., 2011; Soomro et al., 2012) and multi-object tracking (Leal-Taixé et al., 2015), finding that accuracy can degrade up to 77.1% under the most severe corruptions on UCF101.

We then focus on exploring defenses against file/network corruptions, which fall into two broad categories: *corruption-agnostic* methods without the knowledge of bit-level corruptions, and *corruption-aware* methods that exploit knowledge of bit-level corruptions to enforce invariance to file/network corruptions. As corruption-agnostic defenses, we evaluate out-of-distribution (OOD) detection (Hendrycks & Gimpel, 2016; Liang et al., 2017) and adversarial training (Goodfellow et al., 2014). We find that OOD detection can effectively detect corrupted videos at higher corrup-

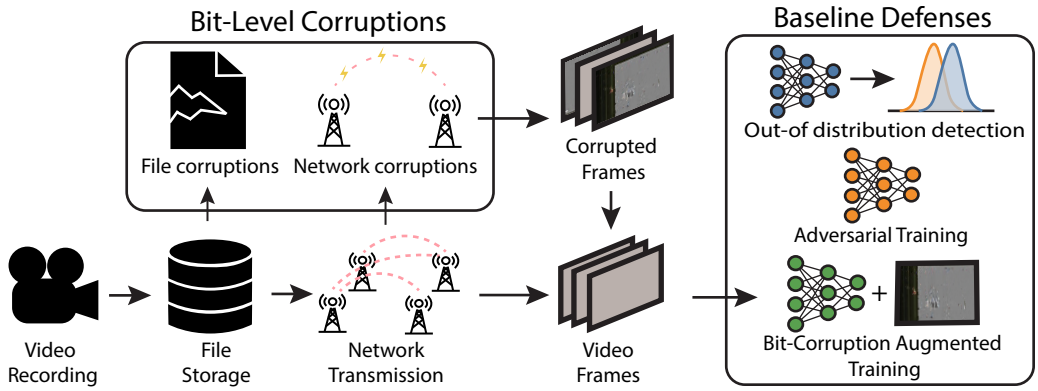


Figure 1: An overview of the video machine learning pipeline under bit-level corruptions, which can arise due to file storage and network transmission. We explore various defenses against file/network corruptions.

tion levels (98.5% AUROC), while adversarial training worsens model performance at almost all levels of corruptions.

Lastly, we propose a corruption-aware training scheme, **B**it-corruption **A**ugmented **T**raining (**BAT**), which augments training data with examples from randomly chosen bit-level corruptions. Our experiments show that BAT can provide up to a 7.1% accuracy boost over the no-defense baseline on highly corrupted videos, but this trades off with accuracy on clean/near-clean data. Further ablation studies suggest that augmenting with severe corruptions is key to increasing robustness.

In summary, our contributions are:

- We investigate the effects of bit-level file and network corruptions, two corruptions previously underexplored in machine learning, on the performance of video ML models.
- We explore corruption-agnostic defense techniques like out-of-distribution detection and adversarial training, finding that existing techniques have limited effectiveness, with OOD detection only being successful at higher corruption levels.
- We propose bit-corruption augmented training (BAT), which increases the robustness of video ML models against bit-level corruptions, and perform ablations by training on targeted ranges of corruptions, improving understanding of corruption-aware defenses.

The rest of the paper is organized as follows. In Section 2, we present our method for simulating file and network corruptions, and defense strategies in Section 3. We present experimental results for our robustness study on bit-level corruptions and defenses in Section 4. Section 5 describes related work, followed by concluding remarks in Section 6.

## 2 Bit-Level Video Corruptions

In this section, we first provide a brief overview on video encoding to provide intuition about the effects of file/network corruptions (Section 2.1), and describe how we simulate file and network corruptions (Section 2.2).

### 2.1 Video Encoding

We present a simplified explanation of video encoding, using the H.264 codec as an illustrative example (Richardson, 2003). Let $\mathcal{X} = \mathbb{R}^D$ be the video space. An uncompressed video $\mathbf{x} \in \mathcal{X}$ consists of a sequence of frames. Content in these frames is often redundant (i.e., two neighboring frames in a video are likely to look similar), so the H.264 codec exploits this by storing a small fraction of frames (known as `I-frames`), and encoding the remaining frames as differences from `I-frames`. To further save space, the frames go through a series of compression steps like frequency space conversion, quantization, and entropy encoding, much like JPEG compression. We notate the encoding function as $\texttt{Enc}(\mathbf{x})$, with a corresponding decoding function $\texttt{Dec}(\cdot)$.

### 2.2 SIMULATING BIT-LEVEL CORRUPTIONS

**Network corruptions.** We study *packet loss*, a network corruption that causes visual quality drops (Hu & Zhang, 2018; Adeyemi-Ejeye et al., 2017). In video streaming, videos are first broken into packets, which store compressed image data or video metadata, then transmitted to a client (Richardson, 2003). Packet loss occurs if the client fails to receive a packet. We simulate this by streaming videos using the Real-Time Streaming Protocol (RTSP) with UDP, dropping video packets randomly (Figure 2). UDP is used in real-time applications like video-conferencing due to latency requirements, but unlike TCP, lost packets in UDP are unrecoverable. Real-world network traces have yielded packet loss rates from 0.31% to 12.38% (Hu & Zhang, 2018).

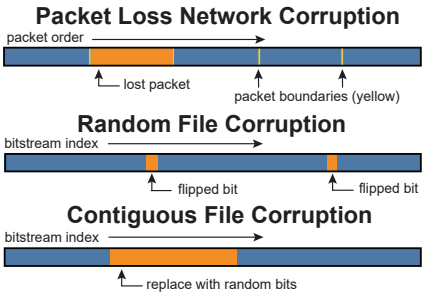


Figure 2: Depiction of bits affected by packet loss network corruption (top), random file corruption (middle), and contiguous file corruption (bottom).

**File corruptions.** We simulate two types of file corruptions, *random corruptions* and *contiguous corruptions* (Figure 2). In random corruptions, random bits in a video file are flipped. Such bit-flips can occur as the result of hardware integrity issues like bus errors (Zhang et al., 2010), malicious software (Sivathanu et al., 2005), or cosmic rays (O'Gorman et al., 1996). In contiguous corruptions, a segment of a video bitstream is replaced with random bits. These errors occur due to sudden temperature changes (thermal asperity), malware (Sivathanu et al., 2005), or misdirected data writes (Zhang et al., 2010). A large-scale study of production flash memory found median bit-error rates across various solid state drive models ranging from $6 \times 10^{-10}$ to $10^{-5}$, with 99th percentile bit error rate reaching $1.2 \times 10^{-4}$ on the least reliable model (Schroeder et al., 2016). Furthermore, despite redundancy mechanisms in modern file systems, many file corruptions are unrecoverable due to unsafe fault-handling or persist undetected (Ganesan et al., 2017).

**Setup.** We describe corruptions with a parameter $p \in [0, 1]$, representing the expected proportion of the video file corrupted. For random corruptions, we flip each bit with probability $p$. To generate contiguous corruptions, we replace a contiguous chunk of length $p$ times the file bitstream length with random bits. For packet loss, each packet is dropped with probability $p$. Since random file corruptions and packet losses are distributed throughout the bitstream, these corruption types affect critical sections of the video at lower $p$. Hence, we simulate random file corruptions and packet loss using a logarithmic scale for $p$. For contiguous corruptions, which are localized, small corruption proportions often do not affect visual quality, so we choose a linear scale for $p$. We provide a probabilistic interpretation of parameter scaling and our exact parameter settings in Appendix D.

## 3 BASELINE DEFENSES FOR BIT-LEVEL CORRUPTIONS

We explore three defenses as baselines for mitigating the impact of file/network corruptions. Broadly, these defenses can be categorized as *corruption-agnostic*, which do not leverage knowledge of file/network corruptions, or *corruption-aware*, which incorporate knowledge of file/network corruptions. We evaluate out-of-distribution detection and adversarial training (AT) as corruption-agnostic defenses, and bit-corruption augmented training (BAT) as a corruption-aware defense.

### 3.1 CORRUPTION-AGNOSTIC DEFENSES

**Out-of-Distribution Detection (OOD).** When deploying a video model trained on clean data, file/network corrupted videos can be viewed as out-of-distribution examples during test time, particularly when the corruption causes large distributional shifts in the visual space. A natural defense for pre-trained video models is OOD detection, which does not require any model retraining.

As an overview of OOD detection, first consider a neural classifier $f(\mathbf{x}; \theta) : \mathbb{R}^D \to \mathbb{R}^K$ with parameters $\theta$, which are optimized under categorical cross-entropy loss:

$$\min_{\theta} \sum_{(\mathbf{x},y)\sim P} \ell(F(\mathbf{x}; \theta), y),$$

where $F(\mathbf{x};\theta)$ is the softmax output. We use ODIN (Liang et al., 2017), a standard OOD detection method that uses calibrated softmax score $S(\mathbf{x};\theta)$ as an indicator score for flagging OOD examples:

$$S(\mathbf{x};\theta) = \max_y F(\mathbf{x};\theta) = \max_y \frac{e^{f_y(\mathbf{x};\theta)/T}}{\sum_i e^{f_i(\mathbf{x};\theta)/T}},$$

where $T$ is the temperature scaling parameter (Guo et al., 2017a) and $F(\mathbf{x};\theta)$ is the post-softmax output. The calibrated score $S(\mathbf{x};\theta)$ is then compared against a threshold; $\mathbf{x}$ is flagged as OOD if it falls below the threshold and is marked as in-distribution otherwise. The idea is that corrupted videos yield lower average calibrated scores than clean videos, making it possible to distinguish clean from corrupted data.

**Adversarial Training (AT).** Adversarial training (Goodfellow et al., 2014) aims to make models robust to imperceptible, adversarial changes to data. Since dversarial noise has small magnitude, we therefore explore whether AT suits visually subtle bit-level corruptions. Adversarial training can be used whenever model retraining is feasible, and does not make any assumptions about bit-level corruptions. This technique optimizes the objective

$$\min_\theta \sum_{(\mathbf{x},y)\sim P} \max_{||\delta||_\infty \leqslant \varepsilon} \ell(F(\mathbf{x}+\delta;\theta), y),$$

where $\varepsilon$ is the perturbation budget, or maximal norm of perturbation $\delta$. We use the training scheme proposed in Wong et al. (2020), where an initial $\delta$ is randomly selected from the $(L_\infty, \varepsilon)$-ball of a video $x$. Then, the above objective is iteratively optimized for one step.

### 3.2 Corruption-Aware Defenses

**Bit-corruption Augmented Training (BAT).** We consider a corruption-aware defense that allow exploits knowledge of bit-level corruptions to enforce invariance to file/network corruptions. In particular, we propose *Bit-corruption Augmented Training* (BAT), which regularizes the video model by augmenting the training data with examples from randomly chosen bit-level corruptions:

$$\min_\theta \sum_{(\mathbf{x},y)\sim P, \tau\sim\mathcal{C}} \ell(F(\texttt{Dec}(\tau(\texttt{Enc}(\mathbf{x}))); \theta), y),$$

where $\mathcal{C}$ is the collection of possible file/network corruptions, and `Enc` and `Dec` denote the video encoding and decoding transformations, respectively. During training, we augment video data by subjecting clean video to one randomly chosen file/network corruption $\tau$ with probability 0.5. This method assumes prior knowledge about file/network corruptions, but not the exact corruption parameters for test-time videos. To better understand the effect of augmenting with different corruptions, we also consider variants of BAT that augment with a subset of corruption parameters as ablations.

## 4 Effect of Bit-level Corruptions on Video Models

We provide an overview of our datasets and models in Section 4.1. In Section 4.2, we study the impact of bit-level corruptions on video model performance. We then investigate the pre-existing corruption-agnostic defenses as baselines (Section 4.3). Finally, in Section 4.4, we experiment with bit-corruption augmented training (BAT) as a corruption-aware defense, finding that this technique improves model robustness to file/network corruptions. We also provide ablations to study the effect of particular sets of corruptions on model robustness.

### 4.1 Datasets and Models

**HMDB51.** The Human Motion DataBase (HMDB51) (Kuehne et al., 2011) is a 51-class action recognition benchmark comprised of movie scenes and web videos. We use a pretrained 3D-Resnet18 architecture fine-tuned on HMDB51 split 1 (3520 train, 1530 test videos), using a standard training setup (Kataoka et al., 2020). At evaluation, we split the input clip into contiguous 16-frame segments, outputting the action class with the highest probability averaged over all segments. We track the accuracy of this pretrained model on a variety of corrupted versions of HMDB51.

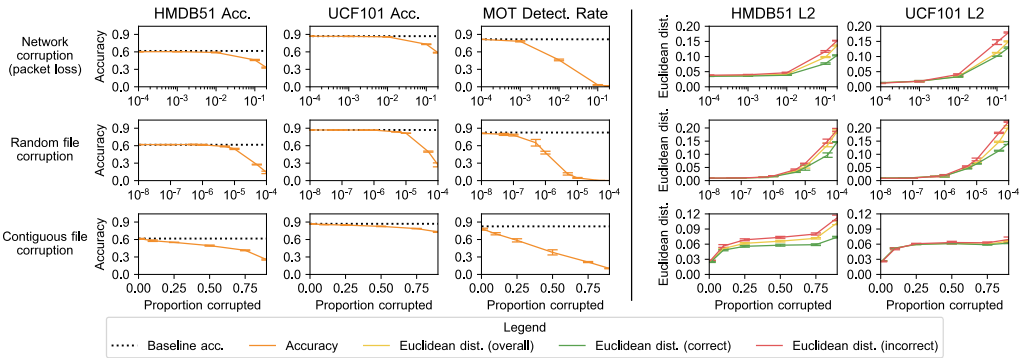


Figure 3: The impact of various file corruption types on model performance (left); average pixel-space Euclidean distance measurements on incorrectly classified examples (red), correctly classified examples (green), and the entire dataset (yellow) organized by corruption proportion and type (right).

**UCF101.** UCF101 (Soomro et al., 2012) is a 101-class action recognition benchmark comprised of web videos. Our setup is identical to that of HMDB51: we fine-tune a pretrained 3D-Resnet18 architecture on UCF101 split 1 (9537 train, 3783 test videos) using the setup of Kataoka et al. (2020). We report accuracy on corrupted versions of UCF101.

**MOT15.** MOT15 (Leal-Taixé et al., 2015) is a multi-object tracking dataset. We use a DLANet-34 (Deep Layer Aggregation) architecture (Zhang et al., 2020) pretrained on MOT20 (Dendorfer et al., 2020). We evaluate on a subset of the *train* split of MOT15 (MOT15-*train*, 6 clips, 12885 objects) following the setup in Zhang et al. (2020). For comparability, we report object detection rate (one minus the false positive rate), which represents the proportion of objects correctly detected, on corrupted versions of MOT15-*train*.

### 4.2 Effects of Bit-Level Corruptions

**Effects of bit-level corruptions on video model performance.** We evaluate model performance on playable[1] videos under bit-level corruptions, finding that model accuracy drops substantially as corruption level increases (Fig. 3). On HMDB51 and UCF101, for random file corruptions, accuracy drops for corruption proportions for $p > 10^{-6}$, and on contiguous file corruptions and network corruptions for $p > 0.01$. On random file corruptions, performance drops by as much as 68.9% on HMDB51 and 77.1% on UCF101; for contiguous file corruptions, performance drops as much as 58.5% on HMDB51 and 16.0% on UCF101. For network corruptions, the drop is 46.5% on HMDB51 and 32.6% on UCF101. For multi-object tracking, we observe that object detection rate drops to near 0 for all types of corruption. Notably, performance declines start at lower corruption levels for this task than for action recognition. More details on multi-object tracking performance are provided in Appendix C.

**Effects of bit-level corruptions in pixel space.** To better understand the effects of bit-level corruptions, we visualize corrupted videos, finding that misclassified videos are more visually distorted than correctly classified videos under corruption. In Fig. 4, under all corruption modes, visual artifacts worsen as corruption level increases. Furthermore, under the same corruption type and proportion, visual artifacts correlate with incorrect model predictions. Fig. 5 shows pairs of clips on each corruption type. We analyze the leftmost pair of clips, which experienced packet loss ($p = 0.1$), but the same general patterns hold on the other clips. Note how packet loss results in noisy patches in the `Kayaking` clip (incorrect), while the sky, grass, and person in the `GolfSwing` example (correct) are not distorted. We observe similar patterns on MOT15, as more distorted videos yield more false positive and false negative object detections. For visualizations of corrupted videos in MOT15, see Appendix C.

We further quantify visual distortion via pixel-space Euclidean distance, calculated as the average $L_2$ norm between pixels in corrupted vs. clean video clips. Fig. 3 shows the average $L_2$ distance between corrupted and clean clips on the entire dataset (yellow), correctly classified examples (green), and incorrectly classified examples (red). On average, incorrect examples have higher $L_2$ distance

[1] Video files can become unplayable if file metadata necessary for decoding is corrupted.

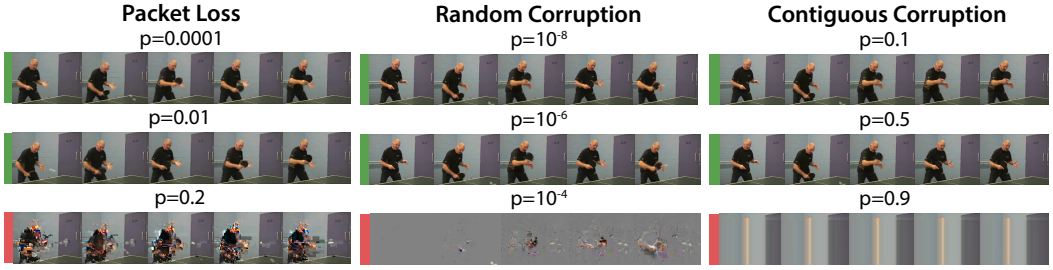


Figure 4: Frames from a clip (class `TableTennisShot`) at varying corruption levels. Color shows correct vs. incorrect classification. Videos become visually more distorted as the corruption level increases.

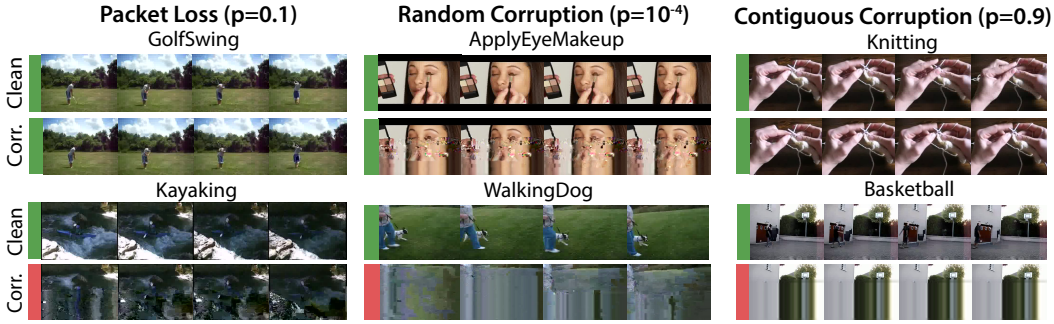


Figure 5: Frames from correctly and incorrectly classified examples, with both clean (1st and 3rd rows) and corrupted (2nd and 4th rows) versions shown. Each pair of corrupted clips were subjected to the **same** corruption proportion $p$ and mode (random file, contiguous file, or packet loss).

than correct examples. For random corruptions, incorrect examples are up to 1.57 more perturbed than correct examples on average (UCF101, random corruption, mid right). Similar observations hold for other corruption types.

### 4.3 Corruption-Agnostic Defenses for Bit-Level Corruptions

**Out-of-distribution Detection.** As observed in Section 4.2, videos with larger pixel-space distortions are more likely to be misclassified, motivating OOD detection as a baseline for filtering out corrupted videos. This defense does not require modifying model parameters, and can be applied to any pre-trained model. Specifically, we use a HMDB51 clean test set as in-distribution data, and a corrupted version of the same split as OOD data. During inference, we apply ODIN (Liang et al., 2017) with $T = 100$ to corrupted datasets as a filter, setting the score threshold to the 5th percentile of calibrated scores on the in-distribution set (95% TPR).

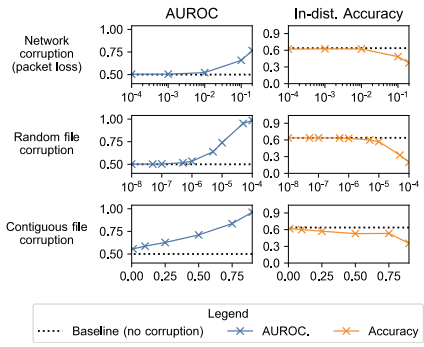


Figure 6: AUROC (left), in-distribution accuracy (right) of OOD detection under packet loss (top), random file corruption (middle), and contiguous file corruption (bottom).

We observe that OOD detection is most successful at higher levels of file corruption. In Figure 6, as corruption proportion increases, in all cases, the AUROC of the OOD detector increases, reaching 98.5% at the highest levels of corruption. OOD detection is less effective at discriminating packet loss, with AUROC only reaching 76.6 at the highest levels of corruption. At low levels of corruption, AUROC degrades to 0.5, as the detector flags many clean examples as OOD. This is expected as corrupted videos become distributionally similar to the clean videos as corruption proportion decreases. Thus, for severely corrupted videos, OOD detection is promising for filtering out corrupted data, but it has trouble discriminating clean from corrupted video at low levels of corruption. Full results can be seen in Appendix B.

Table 1: Robust accuracy on HMDB51 action recognition classification (%) of no defense, adversarial training (AT), and bit-corruption augmented training (BAT) as defenses against network corruptions (top), random file (middle) and contiguous file (bottom).

| Corruption type | $p$ | No defense | **AT** ($\varepsilon = 2/255$) | **BAT** | **BAT** (Low) | **BAT** (High) | **BAT** (Targeted) |
|---|---|---|---|---|---|---|---|
| **Packet loss** | 0.2 | 32.8 | 27.0 | 39.9 | 32.4 | **41.8** | 29.8 |
| | 0.1 | 47.1 | 35.6 | **49.1** | 44.0 | 48.5 | 43.4 |
| | 0.01 | **59.0** | 48.3 | 58.3 | 57.4 | 57.7 | 57.4 |
| | $10^{-3}$ | **60.6** | 49.3 | 58.2 | 57.9 | 58.1 | 57.6 |
| | $10^{-4}$ | **60.0** | 49.4 | 58.9 | 58.4 | 58.7 | 58.0 |
| **Random file corruption** | $10^{-4}$ | 14.1 | 13.2 | 15.0 | 13.0 | **24.2** | |
| | $5 \times 10^{-5}$ | 27.6 | 28.4 | 33.7 | 25.7 | **36.8** | |
| | $10^{-5}$ | 54.5 | 48.0 | 55.6 | 52.8 | **55.7** | |
| | $5 \times 10^{-6}$ | **58.7** | 51.3 | 56.3 | 56.3 | 57.7 | 56.7 |
| | $10^{-6}$ | **61.5** | 53.3 | 58.9 | 58.4 | 59.1 | 57.5 |
| | $5 \times 10^{-7}$ | **62.0** | 53.5 | 59.1 | 58.4 | 59.1 | 57.7 |
| | $10^{-7}$ | **61.9** | 53.5 | 59.1 | 58.5 | 59.0 | 58.5 |
| | $5 \times 10^{-8}$ | **61.9** | 53.5 | 59.1 | 58.4 | 58.9 | 58.6 |
| | $10^{-8}$ | **61.9** | 53.5 | 59.1 | 58.4 | 58.9 | 58.2 |
| **Contiguous file corruption** | 0.9 | 25.6 | 25.8 | 32.2 | 29.2 | 34.9 | **38.1** |
| | 0.75 | 41.4 | 37.2 | 44.1 | 41.2 | **46.0** | 44.6 |
| | 0.5 | 49.5 | 45.0 | 51.6 | 49.6 | **51.7** | 51.6 |
| | 0.25 | **55.0** | 47.8 | 54.9 | 52.9 | 54.6 | 53.9 |
| | 0.1 | **57.9** | 51.0 | 56.3 | 56.8 | 57.4 | 55.1 |
| | 0.01 | **61.2** | 52.0 | 59.5 | 57.8 | 59.2 | 58.5 |
| **None** | 0 | 61.6 | 53.0 | 59.5 | 59.0 | 59.0 | N/A |

**Adversarial Training.** Given the limitations of OOD at lower corruption levels, we explore adversarial training (AT) following the method of Wong et al. (2020), a family of methods for increasing model robustness to small adversarial perturbations (Goodfellow et al., 2014; Madry et al., 2017). The best results ($\varepsilon = 2/255$) are summarized in Table 1; larger $\varepsilon = 4/255, 8/255$ yield worse performance (Appendix A). Interestingly, we find that AT offers little to no utility in increasing robustness to bit-level corruptions, and worsens performance by up to 9.2% accuracy (contiguous corruption, $p = 0.01$). This suggests that the noise structure of bit-level corruptions is fundamentally different from that of adversarial noise. More importantly, this form of AT may not be a viable defense against bit-level corruptions.

### 4.4 Bit-Corruption Augmented Training

**B**it-corruption **A**ugmented **T**raining (BAT) is a corruption-aware defense that augments training data with examples subjected to bit-level file/network corruptions. We report results in Table 1. Across all types of corruption, BAT does slightly worse than a no-defense baseline on clean or near-clean data, which is expected due to the noise in augmented data at high levels of corruption. However, BAT shows significant promise, providing up to **7.1**% lift on highly corrupted data (packet loss, $p = 0.2$) over the no-defense baseline.

To better understand the effect of particular levels of bit-corruption on model robustness, we present ablations where we train two models on two partitions of corruptions: *low* levels of corruption, i.e. those that did not cause a significant performance drop on the baseline model, and *high* levels of corruption, i.e., corruption levels sufficient to cause significant performance drops in the baseline model. Exact corruption parameters for low and high levels of corruption can be found in Appendix D. We find that training on low levels of corruption only (BAT-Low) decreases model robustness across almost all levels of corruptions. In contrast, training with high levels of corruption only (BAT-High) results in an accuracy gain of up to **10.1**% (random file corruption, $p = 10^{-4}$) over the no-defense baseline, further outperforming vanilla BAT at almost all levels of corruption. However, as with vanilla BAT, the lift provided decreases as corruption proportion decreases. This ablation suggests that augmenting with highly corrupted videos in particular is key to increased robustness.

Lastly, we provide an ablation, BAT-targeted, in which we train and test on the same corruption type and level, which makes the strong assumption that we know precisely the corruption parameters the model will encounter. Though unrealistic, BAT-targeted can provide insight into the utility of particular corruptions within the BAT augmentation. Interestingly, this technique lags behind BAT-high at high levels of corruption, and BAT-random, BAT-high, and the no-defense baseline at low levels of corruption. This suggests that training on a variety of different corruptions is important for increasing robustness. Nonetheless, BAT is a promising starting point for robustness to bit-level corruptions in video machine learning, especially for more severe corruptions. Our results motivate future studies of defenses against file/network corruptions to close the gap.

## 5 Related Work

**Robustness in computer vision.** From a non-adversarial perspective, Imagenet-C and -P (Hendrycks & Dietterich, 2019) are ImageNet (Deng et al., 2009) extensions for benchmarking image classifiers on corrupted images. Hendrycks et al. (2020) studies model robustness to distributional shifts like occlusion and perspective shift. On the adversarial side, Szegedy et al. (2013) introduce the concept of adversarial examples, while Goodfellow et al. (2014); Zheng et al. (2016); Madry et al. (2017) extend adversarial ML with methods for creating and defending against adversarial examples. Recent works also extend adversarial attacks to the video domain (Wang & Cherian, 2018; Wei et al., 2018; Inkawhich et al., 2018; Jiang et al., 2019; Naeh et al., 2020). Our work is not only an extension of previous image ML robustness studies to video, but also as a study of file/network corruptions, which remain underexplored in the robustness literature.

**Defenses against corrupted data.** Augmenting data with speckle noise (Rusak et al., 2020), style transfer (Geirhos et al., 2018), smixtures of augmentations (AugMix) (Hendrycks et al., 2019), and distortions constructed from image-to-image networks (DeepAugment) (Hendrycks et al., 2020) have been shown to increase robustness to common corruptions like Imagenet-C. Other defenses include certificate-based methods (Raghunathan et al., 2018a;b) or self-supervision (Carmon et al., 2019). Non-adversarial perturbations can be interpreted as OOD data, for which supervised (Lee et al., 2018; Bahat & Shakhnarovich, 2018; Geifman & El-Yaniv, 2019; Gorbett & Blanchard, 2020) and semi-supervised OOD detection techniques (Liu et al., 2018) have been proposed. Video anomaly detection techniques are also explored in (Ben-Ari & Shwartz-Ziv, 2018; Gutoski et al., 2017). From an adversarial perspective, common defenses include adversarial training (Goodfellow et al., 2014; Madry et al., 2017). We evaluate baseline non-adversarial and adversarial robustness techniques, and extend the augmentation-as-defense paradigm to bit-level corruptions.

**Bit-space aware robustness studies.** Previous work has studied the robustness of computer vision models under various aspects of image and video compression, such as compression rates or encoding schemes (Srinivasan et al., 2016; Seymour et al., 2007). On the network side, deep learning-based packet loss concealment (Lee & Chang, 2016; Lotfidereshgi & Gournay, 2018; Mohamed & Schuller, 2020) techniques aim to recover from packet loss by predicting lost information. We view our work as complementary to previous efforts, extending robustness studies to explicitly examine the effect of file/network corruptions at the bit-level on video model performance.

## 6 Conclusion

In this paper, we take a first step in investigating the effect of file and network corruptions on video model robustness, finding that these corruptions significantly drop performance on benchmark tasks like action recognition and multi-object tracking. We try corruption-agnostic defenses against file/network corruptions, finding that out-of-distribution detection shows promise at high corruption levels, and that adversarial training is ineffective. We also evaluate our method of bit-corruption augmented training (BAT), a corruption-aware defense, which improves performance at high corruption levels, making it a promising starting point for future studies of file/network corruptions and model robustness. The problem of robustness in video ML to file and network corruptions remains unsolved, so we hope this work motivates future studies of defenses against these corruptions.

## REFERENCES

A.O. Adeyemi-Ejeye, M. Alreshoodi, L. Al-Jobouri, M. Fleury, and J. Woods. Packet loss visibility across sd, hd, 3d, and uhd video streams. *Journal of Visual Communication and Image Representation*, 45:95 – 106, 2017. ISSN 1047-3203. doi: https://doi.org/10.1016/j.jvcir.2017.02.012. URL http://www.sciencedirect.com/science/article/pii/S1047320317300494.

Ayse Elvan Aydemir, Alptekin Temizel, and Tugba Taskaya Temizel. The effects of jpeg and jpeg2000 compression on attacks using adversarial examples, 2018.

Yuval Bahat and Gregory Shakhnarovich. Confidence from invariance to image transformations, 2018.

Itamar Ben-Ari and Ravid Shwartz-Ziv. Attentioned convolutional LSTM inpaintingnetwork for anomaly detection in videos. *CoRR*, abs/1811.10228, 2018. URL http://arxiv.org/abs/1811.10228.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks, 2016.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C. Duchi. Unlabeled data improves adversarial robustness, 2019.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2017.

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset, 2019.

Yiping Chen, Jingkang Wang, Jonathan Li, Cewu Lu, Zhipeng Luo, Han Xue, and Cheng Wang. Lidar-video driving dataset: Learning driving policies effectively. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5870–5878, 2018.

Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E. Kounavis, and Duen Horng Chau. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression, 2018.

Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes, 2020.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M. Roy. A study of the effect of jpg compression on adversarial images, 2016.

Daniel Y. Fu, Will Crichton, James Hong, Xinwei Yao, Haotian Zhang, Anh Truong, Avanika Narayan, Maneesh Agrawala, Christopher Ré, and Kayvon Fatahalian. Rekall: Specifying video events using compositions of spatiotemporal labels. *arXiv preprint arXiv:1910.02993*, 2019.

Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. Temporal localization of actions with actoms. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2782–2795, 2013.

Aishwarya Ganesan, Ramnatthan Alagappan, Andrea C. Arpaci-Dusseau, and Remzi H. Arpaci-Dusseau. Redundancy does not imply fault tolerance: Analysis of distributed storage reactions to file-system faults. *ACM Trans. Storage*, 13(3), September 2017. ISSN 1553-3077. doi: 10.1145/3125497. URL https://doi.org/10.1145/3125497.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option, 2019.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2018.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014.

Matt Gorbett and Nathaniel Blanchard. Utilizing network properties to detect erroneous inputs, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1321–1330. JMLR. org, 2017a.

Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. *CoRR*, abs/1711.00117, 2017b. URL http://arxiv.org/abs/1711.00117.

Matheus Gutoski, Nelson Marcelo Romero Aquino, Manassés Ribeiro, André Lazzaretti, and Heitor Lopes. Detection of video anomalies using convolutional autoencoders and one-class support vector machines. 11 2017. doi: 10.21528/CBIC2017-49.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks, 2016.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization, 2020.

Zhiguo Hu and Qiqiang Zhang. A new approach for packet loss measurement of video streaming and its application. *Multimedia Tools Appl.*, 77(10):11589–11608, May 2018. ISSN 1380-7501. doi: 10.1007/s11042-016-3566-0. URL https://doi.org/10.1007/s11042-016-3566-0.

Nathan Inkawhich, Matthew Inkawhich, Yiran Chen, and Hai Li. Adversarial attacks for optical flow-based action recognition classifiers, 2018.

Bart Jansen, Timothy Goodwin, Varun Gupta, Fernando Kuipers, and Gil Zussman. Performance evaluation of webrtc-based video conferencing. *SIGMETRICS Perform. Eval. Rev.*, 45(3):56–68, March 2018. ISSN 0163-5999. doi: 10.1145/3199524.3199534. URL https://doi.org/10.1145/3199524.3199534.

Linxi Jiang, Xingjun Ma, Shaoxiang Chen, James Bailey, and Yu-Gang Jiang. Black-box adversarial attacks on video recognition models. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 864–872, 2019.

Hirokatsu Kataoka, Tenga Wakamiya, Kensho Hara, and Yutaka Satoh. Would mega-scale datasets further enhance spatiotemporal 3d cnns?, 2020.

H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking, 2015.

B. Lee and J. Chang. Packet loss concealment based on deep neural networks for digital speech transmission. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2):378–387, 2016.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks, 2018.

Shiyu Liang, Yixuan Li, and R. Srikant. Principled detection of out-of-distribution examples in neural networks. *CoRR*, abs/1706.02690, 2017. URL http://arxiv.org/abs/1706.02690.

Si Liu, Risheek Garrepalli, Thomas G. Dietterich, Alan Fern, and Dan Hendrycks. Open category detection with pac guarantees, 2018.

R. Lotfidereshgi and P. Gournay. Speech prediction using an adaptive recurrent neural network with application to packet loss concealment. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5394–5398, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2017.

Mostafa M. Mohamed and Björn W. Schuller. Concealnet: An end-to-end neural network for packet loss concealment in deep speech emotion recognition, 2020.

Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, and Aude Oliva. Moments in time dataset: one million videos for event understanding, 2018.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations, 2016.

Muhammad-Sajid Mushtaq and Abdelhamid Mellouk. 2 - methodologies for subjective video streaming qoe assessment. In Muhammad-Sajid Mushtaq and Abdelhamid Mellouk (eds.), *Quality of Experience Paradigm in Multimedia Services*, pp. 27 – 57. Elsevier, 2017. ISBN 978-1-78548-109-3. doi: https://doi.org/10.1016/B978-1-78548-109-3.50002-3. URL http://www.sciencedirect.com/science/article/pii/B9781785481093500023.

Itay Naeh, Roi Pony, and Shie Mannor. Flickering adversarial attacks against video recognition networks, 2020.

Ravi Netravali, Anirudh Sivaraman, Somak Das, Ameesh Goyal, Keith Winstein, James Mickens, and Hari Balakrishnan. Mahimahi: Accurate record-and-replay for HTTP. In *2015 USENIX Annual Technical Conference (USENIX ATC 15)*, pp. 417–429, Santa Clara, CA, July 2015. USENIX Association. ISBN 978-1-931971-225. URL https://www.usenix.org/conference/atc15/technical-session/presentation/netravali.

T. J. O'Gorman, J. M. Ross, A. H. Taber, J. F. Ziegler, H. P. Muhlfeld, C. J. Montrose, H. W. Curtis, and J. L. Walsh. Field testing for cosmic ray soft errors in semiconductor memories. *IBM Journal of Research and Development*, 40(1):41–50, 1996.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples, 2018a.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Semidefinite relaxations for certifying robustness to adversarial examples, 2018b.

Iain E. G. Richardson. *H.264 and MPEG-4 Video Compression*. Wiley, 2003.

Evgenia Rusak, Lukas Schott, Roland S. Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A simple way to make neural networks robust against diverse image corruptions, 2020.

Bianca Schroeder, Raghav Lagisetty, and Arif Merchant. Flash reliability in production: The expected and the unexpected. In *14th USENIX Conference on File and Storage Technologies (FAST 16)*, pp. 67–80, Santa Clara, CA, February 2016. USENIX Association. ISBN 978-1-931971-28-7. URL https://www.usenix.org/conference/fast16/technical-sessions/presentation/schroeder.

Rowan Seymour, Darryl Stewart, and Ji Ming. Comparison of image transform-based features for visual speech recognition in clean and corrupted videos. *EURASIP Journal on Image and Video Processing*, 2007. doi: doi:10.1155/2008/810362.

Gopalan Sivathanu, Charles P. Wright, and Erez Zadok. Ensuring data integrity in storage: Techniques and applications. In *Proceedings of the 2005 ACM Workshop on Storage Security and Survivability*, pp. 26–36, 2005.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012.

V. Srinivasan, S. Gul, S. Bosse, J. T. Meyer, T. Schierl, C. Hellge, and W. Samek. On the robustness of action recognition methods in compressed and pixel domain. In *2016 6th European Workshop on Visual Information Processing (EUVIP)*, pp. 1–6, 2016.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2013.

Jue Wang and Anoop Cherian. Contrastive video representation learning via adversarial perturbations, 2018.

Xingxing Wei, Jun Zhu, and Hang Su. Sparse adversarial perturbations for videos, 2018.

Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training, 2020.

Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. A simple baseline for multi-object tracking, 2020.

Yupu Zhang, Abhishek Rajimwale, Andrea C. Arpaci-Dusseau, and Remzi H. Arpaci-Dusseau. End-to-end data integrity for file systems: A zfs case study. In *Proceedings of the 8th USENIX Conference on File and Storage Technologies*, pp. 3, 2010.

Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.