Do Image Classifiers Generalize Across Time?

Vaishaal Shankar*
UC Berkeley

Achal Dave*
CMU

Rebecca Roelofs
UC Berkeley

Deva Ramanan
CMU

Benjamin Recht
UC Berkeley

Ludwig Schmidt
UC Berkeley

Abstract

Vision models notoriously flicker when applied to videos: they correctly recognize objects in some frames, but fail on perceptually similar, nearby frames. In this work, we systematically analyze the robustness of image classifiers to such temporal perturbations in videos. To do so, we construct two new datasets, `ImageNet-Vid-Robust` and `YTBB-Robust`, containing a total of 57,897 images grouped into 3,139 sets of perceptually similar images. Our datasets were derived from ImageNet-Vid and Youtube-BB, respectively, and thoroughly re-annotated by human experts for image similarity. We evaluate a diverse array of classifiers pre-trained on ImageNet and show a median classification accuracy drop of 16 and 10 points, respectively, on our two datasets. Additionally, we evaluate three detection models and show that natural perturbations induce both classification as well as localization errors, leading to a median drop in detection mAP of 14 points. Our analysis demonstrates that perturbations occurring naturally in videos pose a substantial and realistic challenge to deploying convolutional neural networks in environments that require both reliable and low-latency predictions.

1. Introduction

Applying state-of-the-art image recognition systems to videos reveals a troubling phenomenon: models correctly recognize objects in one frame, but fail to do so in the very next frame (Figure 1). In practice, this *flickering* of predictions is treated as an unfortunate but unavoidable property of image-based models. This issue can be mitigated in offline settings by smoothing predictions over time. However, online smoothing isn't nearly as effective and incurs a delay, resulting in catastrophic mistakes in downstream applications: *e.g.*, flickering object classifications have reportedly led to fatal autonomous vehicle collisions [3].

At its root, prediction flicker is a manifestation of a broader issue: current models lack *robustness* to small input perturbations. In the machine learning community, model robustness has typically been analyzed on images perturbed by an adversary [11, 2], or by hand-designed strategies, such as rotations or blurs [7, 6, 14, 13]. However, these benchmarks rely on synthetically modifying the input image, serving at best as *proxies* for evaluating robustness to *natural* perturbations, which are common in videos.

In this work, we systematically analyze the prevalence of flicker across vision models. Taking inspiration from the

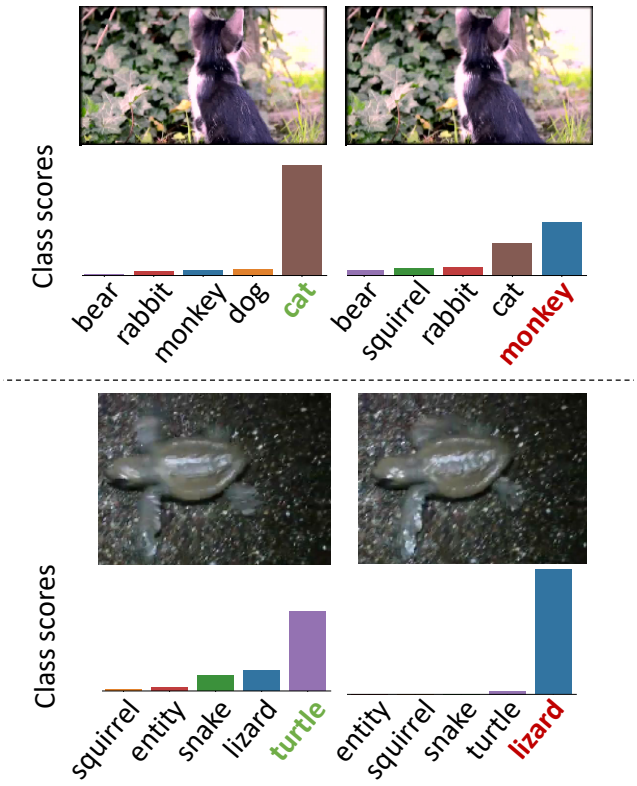

Figure 1: Examples of natural perturbations from nearby video frames and resulting classifier predictions from a ResNet-152 model fine-tuned on ImageNet-Vid. While the images appear almost identical to the human eye, the classifier confidence changes substantially.

*Equal contribution

robustness literature, we evaluate models on *perceptually similar* images, which we sample from nearby video frames. However, nearby frames can still exhibit drastic changes (*e.g.*, significant occlusions), which may cause even robust models to fail. We discard such frame pairs by employing human expert labelers to evaluate model robustness only on perceptually similar images, unlike prior work [12]. As a cornerstone of our investigation, we introduce two test sets for evaluating model robustness: `ImageNet-Vid-Robust` and `YTBB-Robust`, carefully curated from the ImageNet-Vid and Youtube-BB datasets [27, 24]. To the best of our knowledge these are the first datasets of their kind, containing tens of thousands of images that are *human reviewed* and grouped into thousands of perceptually similar sets. In total, our datasets contain 3,139 sets of temporally adjacent and visually similar images (57,897 images total).

We use these datasets to measure the robustness of current models to small, naturally occurring perturbations. Although we use videos to sample these images, our datasets allow evaluating the robustness of standard, image-based computer vision models, such as those trained on ImageNet. Our testbed contains over 47 different models, varying model types (CNNs, transformers), architectures (*e.g.*, AlexNet, ResNet) and training methods (*e.g.*, adversarial training, augmentation). To systematically characterize flicker, we also introduce a stringent robustness metric.

Our experiments show that all models in our testbed degrade significantly in the presence of small, natural perturbations in video frames. Under our metric, we find such perturbations in `ImageNet-Vid-Robust` and `YTBB-Robust` induce median accuracy drops of 16% and 10% respectively for classification, and a median 14 point AP drop for detection[1]. Even for the best-performing classification models trained on public datasets, we observe an accuracy drop of 14% for `ImageNet-Vid-Robust` and 8% for `YTBB-Robust`. Recently introduced, contrastive models trained on weakly supervised web images [23] can reduce this gap, but require over 400 million images, and still exhibit noticeable gaps of 6.1% and 6.7%, respectively.

Our results show that robustness to natural perturbations in videos is problematic for a wide variety of models. Practical deployment of models, especially in safety-critical environments like autonomous driving, requires predictions that are not only accurate, but also robust over time. Our analysis indicates that ensuring reliable predictions on *every frame* of a video is an important direction for future work.

2. Related work

Adversarial examples. While various forms of adversarial examples have been studied, the majority of research focuses on ℓ_p robustness [11, 2, 32]. However, it is unclear whether adversarial examples pose a problem for robustness outside of a truly worst case context. It is an open question whether perfect robustness against a ℓ_p adversary will induce robustness to realistic image distortions such as those studied in this paper. Recent work has proposed less adversarial image modifications such as small rotations & translations [6, 1, 7, 17], hue and color changes [14], image stylization [9] and synthetic image corruptions such as Gaussian blur and JPEG compression [13, 10]. Even though the above examples are more realistic than the ℓ_p model, they still synthetically modify the input images to generate perturbed versions. In contrast, our work performs no synthetic modification and instead uses unmodified video frames.

Studying robustness in videos. In recent work, [12] exploit the temporal structure in videos to study robustness. However, their experiments suggest a substantially smaller drop in accuracy. The primary reason for this is a less stringent metric used in [12]. By contrast, our PM-k metric is inspired by the "worst-of-k" metric used in prior work [6], highlighting the sensitivity of models to natural perturbations. In the appendix, we study the differences between the two metrics in more detail. Furthermore, the lack of human review and the high label error-rate we discovered in Youtube-BB (Table 1) presents a potentially troubling confounding factor that we resolve in our work.

Distribution shift. Small, benign changes in the test distribution are often referred to as *distribution shift*. [25] explore this phenomenon by constructing new test sets for CIFAR-10 and ImageNet and observe substantial performance drops for a large suite of models on the newly constructed test sets. Similar to our Figure 3, the relationship between their original and new test set accuracies is also approximately linear. However, the images in their test set bear little visual similarity to images in the original test set, while all of our failure cases are on perceptually similar images. In a similar vein of study, [29] studies distribution shift *across* different computer vision data sets such as Caltech-101, PASCAL, and ImageNet.

Temporal consistency in computer vision. Authors of [16] explicitly identify flickering failures and use a technique reminiscent of adversarially robust training to improve image-based models. A similar line of work focuses on improving object detection in videos as objects become occluded or move quickly [18, 8, 33, 30]. The focus in this work has generally been on improving object detection when objects transform in a way that makes recognition difficult from a single frame, such as fast motion or occlusion. In this work, we document a broader set of failure cases for image-based classifiers and detectors and show that failures occur when the neighboring frames are imperceptibly different.

[1] We only evaluated detection on `ImageNet-Vid-Robust` as bounding-box labels in Youtube-BB are not temporally dense enough for our evaluation.

Figure 2: Temporally adjacent frames may not be visually similar. We show three randomly sampled frame pairs where the nearby frame was marked as "dissimilar" to the anchor frame during human review and then discarded from our dataset.

		ImageNet-Vid Robust		YTBB Robust	
Anchor frames	Reviewed	1,314		2,467	
	Accepted	1,109	(84%)	2,030	(82%)
	Labels updated	-		834	(41%)
Frame pairs	Reviewed	26,029		45,631	
	Accepted	21,070	(81%)	36,827	(81%)

Table 1: Dataset statistics of `ImageNet-Vid-Robust` and `YTBB-Robust`. For `YTBB-Robust`, we updated the labels from for 41% (834) of the accepted anchors due to incomplete labels in Youtube-BB.

3. Evaluating temporal robustness

`ImageNet-Vid-Robust` and `YTBB-Robust` are sourced from videos in the ImageNet-Vid and Youtube-BB datasets [27, 24]. All but one [2] of the object classes in ImageNet-Vid and Youtube-BB are from the WordNet hierarchy [21] and direct ancestors of ILSVRC-2012 classes. Using the WordNet hierarchy, we construct a canonical mapping from ILSVRC-2012 classes to ImageNet-Vid and Youtube-BB classes, which allows us to evaluate off-the-shelf ILSVRC-2012 models on `ImageNet-Vid-Robust` and `YTBB-Robust`. We provide more background on the source datasets in the appendix.

3.1. Dataset construction

Next, we describe how we extracted sets of naturally perturbed frames from ImageNet-Vid and Youtube-BB to create `ImageNet-Vid-Robust` and `YTBB-Robust`. A straightforward approach would be to select a set of anchor frames and use temporally adjacent frames in the video with the assumption that such frames contain only small perturbations from the anchor. However, as Figure 2 illustrates, this assumption is frequently violated, especially due to fast camera or object motion.

Instead, we first collect *preliminary* datasets of natural perturbations following the same approach, and then manually review each of the frame sets. For each video, we randomly sample an anchor frame and take $k = 10$ frames before and after the anchor frame as candidate perturbation images[3]. This results in two datasets containing one anchor frame each from 3,139 videos, with approximately 20 candidate perturbation per anchor frame[4].

Next, we curate the dataset with the help of four expert human annotators. The goal of the curation step is to ensure that each anchor frame and its nearby frames are correctly labeled with the same ground truth class, and that the anchor frame and the nearby frames are visually similar.

Denser labels for Youtube-BB. As Youtube-BB contains only a single category label per frame at 1 frame per second, annotators first inspected each anchor frame individually and added any missing labels. In total, annotators corrected the labels for 834 frames, adding an average of 0.5 labels per anchor frame. These labels are then propagated to nearby, unlabeled frames at the native frame rate and verified in the next step. ImageNet-Vid densely labels all classes per frame, so we skipped this step for this dataset.

Frame pairs review. Next, for each pair of anchor and nearby frames, a human annotates (i) whether the pair is correctly labeled in the dataset, and (ii) whether the pair is similar. We took several steps to mitigate the subjectivity of this task and ensure high annotation quality. First, we trained reviewers to mark frames as dissimilar if the scene undergoes any of the following transformations: significant motion, significant background change, or significant blur change. We asked reviewers to mark each dissimilar frame with one of these transformations, or "other", and to mark a pair of images as dissimilar if a distinctive feature of the object is only visible in one of the two frames (such as the face of a dog). If an annotator was unsure about the correct label, she could mark the pair as "unsure". Second, we present only a single pair of frames at a time to reviewers because presenting videos or groups of frames could cause them to miss large changes due to the phenomenon of change blindness [22].

[2] the class "skateboard" in Youtube-BB is not present in ILSVRC-2012

[3] For `YTBB-Robust` we use a subset of the anchor frames used by [12].

[4] Anchor frames near the start or end of the video may have less than 20 candidate frames.

Verification. In the previous stage, all annotators were given identical labeling instructions and individually reviewed a total of 71,660 image pairs. To increase consistency in annotation, annotators jointly reviewed all frames marked as dissimilar, incorrectly labeled, or "unsure". A frame was only considered similar to its anchor if a strict majority of the annotators marked the pair as such.

After the reviewing was complete, we discarded all anchor frames and candidate perturbations that annotators marked as dissimilar or incorrectly labeled. The final datasets contain a combined total of 3,139 anchor frames with a median of 20 similar frames each.

3.2. The `pm-k` evaluation metric

Given the datasets introduced above, we propose a metric to measure a model's robustness to natural perturbations. In particular, let $A = \{a_1, ..., a_n\}$ be the set of valid anchor frames in our dataset. Let $Y = \{y_1, ..., y_n\}$ be the set of labels for A. We let $\mathcal{N}_k(a_i)$ be the set of frames marked as similar to anchor frame a_i. In our setting, $\mathcal{N}_k$ is a subset of the $2k$ temporally adjacent frames (plus/minus k frames from the anchor).

Classification. The standard classification accuracy on the anchor frame is $\text{acc}_{\text{orig}} = 1 - \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{0/1}(f(a_i), y_i)$, where $\mathcal{L}_{0/1}$ is the standard 0-1 loss function. We define the `pm-k` analog of accuracy as

$$\text{acc}_{\text{pmk}} = 1 - \frac{1}{N}\sum_{i=1}^{N} \max_{b \in \mathcal{N}_k(a_i)} \mathcal{L}_{0/1}(f(b), y_i) \ , \quad (1)$$

which corresponds to picking the worst frame from each set $\mathcal{N}_k(a_i)$ before computing accuracy. We note the similarity of the `pm-k` metric to standard ℓ_p-robustness. If we let $\mathcal{N}_k(a_i)$ be the set of *all* images within an ℓ_p ball of radius ϵ around a_i, then the notions of robustness are identical. For frames with multiple labels, we count a prediction as correct if the model predicts *any* of the correct classes for a frame for both accuracy measures.

Detection. The standard metric for detection is mean average precision (mAP) of the predictions at a fixed intersection-over-union (IoU) threshold [19]. We define the `pm-k` metric analogous to that for classification: We replace each anchor frame with the nearest frame that minimizes the average precision (AP, averaged over recall thresholds) of the predictions, and compute `pm-k` as the mAP on these worst-case neighboring frames.

4. Main results

We evaluate a testbed of 47 classification and three detection models on `ImageNet-Vid-Robust` and `YTBB-Robust`. We first discuss the various types of classification models evaluated with the `pm-k` classification metric. Second, we evaluate the performance of detection models on `ImageNet-Vid-Robust` using use the bounding box annotations inherited from ImageNet-Vid and using a variant of the `pm-k` metric for detection. We then analyze the errors made on the detection adversarial examples to isolate the effects of *localization* errors vs. *classification* errors. Finally, we analyze the impact of dataset review, video compression, and video frame rate on the accuracy drop.

4.1. Classification

The classification robustness metric is acc_{pmk} defined in Equation (1). In Figure 3, we plot the benign accuracy, acc_{orig}, versus the robust accuracy, acc_{pmk}, for all classification models in our test bed and find a consistent drop from acc_{orig} to acc_{pmk}. Further, we note that the relationship between acc_{orig} and acc_{pmk} is approximately linear, indicating that while improvements in the benign accuracy do result in improvements in the worst-case accuracy, they do not suffice to resolve the accuracy drop due to natural perturbations. We provide implementation details and hyperparameters for all models in the supplementary.

Our test bed consists of six model types with increasing levels of supervision. We present results for representative models from each model type in Section 4.1.

ILSVRC Trained The WordNet hierarchy enables us to repurpose models trained for the 1,000 class ILSVRC-2012 dataset on `ImageNet-Vid-Robust` and `YTBB-Robust` We evaluate a wide array of ILSVRC-2012 models (available from [4]) against our natural perturbations. Since these datasets present a substantial distribution shift from the original ILSVRC-2012 validation set, we expect the *benign* accuracy acc_{orig} to be lower than the comparable accuracy on the ILSVRC-2012 validation set. However, our main interest here is in the *difference* between the original and perturbed accuracies acc_{orig} - acc_{pmk}. A small drop in accuracy would indicate that the model is robust to small changes that occur naturally in videos. Instead, we find significant median drops of 15.0% and 13.2% in accuracy on our two datasets, indicating sensitivity to such changes.

Noise augmentation One hypothesis for the accuracy drop from original to perturbed accuracy is that subtle artifacts and corruptions introduced by video compression schemes could degrade performance when evaluating on these corrupted frames. The worst-case nature of the `pm-k` metric could then be focusing on these corrupted frames. One model for these corruptions are the perturbations introduced in [13]. To test this hypothesis, we evaluate models augmented with a subset of the perturbations (exactly one of: Gaussian noise, Gaussian blur, shot noise, contrast change,

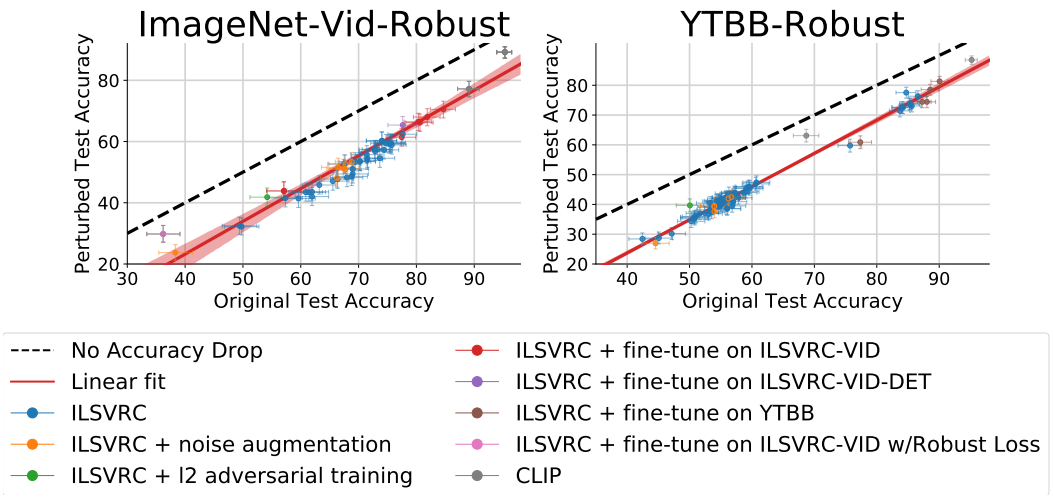

Figure 3: Model accuracy on original vs. perturbed images. Each data point corresponds to one model in our testbed (shown with 95% Clopper-Pearson confidence intervals). If models were robust to perturbations, we would expect them to fall on the dashed line ($y = x$). Instead, we find they all lie significantly below this ideal line, consistently exhibiting a significant accuracy drop to perturbed frames. Each perturbed frame was taken from a ten frame neighborhood (approximately 0.3 seconds) of the original frame, and reviewed by experts to confirm visual similarity to the original frame.

Model Type	Accuracy Original	Accuracy Perturbed	Δ
`ImageNet-Vid-Robust`			
Trained on ILSVRC	67.5 [64.7, 70.3]	52.5 [49.5, 55.5]	15.0
+ Noise Augmentation	68.8 [66.0, 71.5]	53.2 [50.2, 56.2]	15.6
+ ℓ_∞ robustness (ResNext-101)	54.3 [51.3, 57.2]	40.8 [39.0, 43.7]	12.4
+ FT on ImageNet-Vid	80.8 [78.3, 83.1]	65.7 [62.9, 68.5]	15.1
+ FT PM-k loss on ImageNet-Vid	36.2 [33.3, 39.1]	29.8 [27.1, 32.5]	6.4
+ FT on ImageNet-Vid (ResNet-152)	84.8 [82.5, 86.8]	70.2 [67.4, 72.8]	14.6
+ FT on ImageNet-Vid-Det	77.6 [75.1, 80.0]	65.4 [62.5, 68.1]	12.3
CLIP Zero-Shot	95.3 [93.8, 96.4]	89.2 [87.2, 91.0]	6.1
`YTBB-Robust`			
Trained on ILSVRC	57.0 [54.9, 59.2]	43.8 [41.7, 46.0]	13.2
+ Noise Augmentation	62.3 [60.2, 64.4]	45.7 [43.5, 47.9]	16.6
+ ℓ_∞ robustness (ResNext-101)	53.6 [51.4, 55.8]	43.2 [41.0, 45.3]	10.4
+ FT on Youtube-BB	91.4 [90.1, 92.6]	82.0 [80.3, 83.7]	9.4
+ FT on Youtube-BB (ResNet-152)	92.9 [91.6, 93.9]	84.7 [83.0, 86.2]	8.2
CLIP Zero-Shot	95.2 [93.9, 95.8]	88.5 [87.0, 89.8]	6.7

Table 2: Accuracies of six model types and the best performing model (shown with 95% Clopper-Pearson confidence intervals). Δ denotes accuracy drop between evaluation on anchor frame (acc_{orig}) and worst frame in similarity set (acc_{pmk}). The model architecture is ResNet-50 unless noted otherwise. 'FT' denotes 'fine-tuning.' See Section 4.1 for details.

impulse noise, or JPEG compression). We found that these augmentation schemes did not improve robustness against our perturbations substantially, and still result in a median accuracy drop of 15.6% and 16.6% on the two datasets.

ℓ_∞-robustness. We evaluate the model from [31], which currently performs best against ℓ_∞-attacks on ImageNet. We find that this model has a smaller accuracy drop than the two aforementioned model types on both datasets. However, the robust model achieves substantially lower original and perturbed accuracy than either of the two model types above, and the robustness gain is modest (3% compared to models of similar benign accuracy). In section 4.3 of [28], the authors further analyze the performance of ℓ_∞-robust models on

Task	Model	mAP Original	mAP Perturbed	mAP Δ
Detection	FRCNN, ResNet 50	62.8	48.8	14.0
	FRCNN, ResNet 101	63.1	50.6	12.5
	R-FCN, ResNet 101 [30]*	79.4*	63.7*	15.7*
Localization	FRCNN, ResNet 50	76.6	64.2	12.4
	FRCNN, ResNet 101	77.8	66.3	11.5
	R-FCN, ResNet 101*	80.9*	70.3*	10.6*

Table 3: Detection and localization mAP for Faster R-CNN and R-FCN models. Both detection and localization suffer from significant mAP drops due to perturbations. (R-FCN was trained on ILSVRC Det and VID 2015, and evaluated on the 2015 subset of ILSVRC-VID 2017, indicated by *.)

Figure 4: Naturally perturbed examples for detection. Red boxes indicate false positives; green boxes indicate true positives; white boxes are ground truth. Classification errors are common failures, such as the fox on the left, which is classified correctly in the anchor frame, and misclassified as a sheep in a nearby frame. However, detection models also have *localization* errors, where the object of interest is not correctly localized in addition to being misclassified, such as the airplane (middle) and the motorcycle (right). All visualizations show predictions with confidence greater than 0.5.

`ImageNet-Vid-Robust` and `YTBB-Robust`.

Fine-tuning on video frames. To adapt to the new class vocabulary and the video domain, we fine-tune several network architectures on the ImageNet-Vid and Youtube-BB training sets. For Youtube-BB, we train on the anchor frames used for training in [12], and for ImageNet-Vid we use all frames in the training set. The resulting models significantly improve in accuracy over their ILSVRC pre-trained counterparts (e.g., 13% on `ImageNet-Vid-Robust` and 34% on `YTBB-Robust` for ResNet-50). This improvement in accuracy results in a modest improvement in robustness for `YTBB-Robust`, but still suffers from a substantial 9.4% drop. On `ImageNet-Vid-Robust`, there is almost no change in the drop from 15.0% to 15.1%.

Fine-tuning with a robust loss. Training on videos optimizes for the *average* accuracy on video frames. However, our goal at test-time is to improve the worst-case, PM-k accuracy. We adopt a strategy inspired by work in adversarial robustness [20], which uses the PM-k metric as the training loss. Specifically, for each frame x_t, let the standard training loss be for a model f be $L(x_t, y_t; f)$. We instead train the model using

$$\hat{L}(f(x_t), y_t) = \max_{\hat{x} \in \mathcal{N}_k(x_t)} L(f(\hat{x}), y_t),$$

where $\mathcal{N}_k(x_t)$ contains all images within k frames of x_t with labels that match y_t. Unfortunately, this results in a drastic drop in both the original and perturbed accuracies by 31.3% and 22.7% respectively. However, the strategy does reduce the robustness gap from 15.1% to 6.4%, suggesting this loss may be a promising avenue for future improvements in robustness. We provide implementation details and further

	Reviewed	Accuracy		
		Original	Perturbed	Δ
ImageNet-Vid-Robust	✗	80.3	64.1	16.2
	✓	84.8	70.2	14.4
YTBB-Robust	✗	88.1	78.1	10.0
	✓	92.9	84.7	8.9

Table 4: Impact of human review on ImageNet-Vid-Robust and YTBB-Robust on original and perturbed accuracy, using ResNet-152 fine-tuned on ImageNet-Vid and Youtube-BB, respectively.

analysis of this model in the supplementary.

Fine-tuning for detection on video frames. We further analyze whether additional supervision in the form of bounding box annotations improves robustness. To this end, we train the Faster R-CNN *detection* model [26] with a ResNet-50 backbone on ImageNet-Vid. Following standard practice, the detection backbone is pre-trained on ILSVRC-2012. To evaluate this detector for classification, we assign the class with the most confident bounding box as label to the image. We find that this transformation reduces accuracy compared to the model trained for classification (77.6% vs. 80.8%). While there is a slight reduction in the accuracy drop caused by natural perturbations, the reduction is well within the error bars for this test set.

Contrastive Language-Image Pre-training (CLIP) Recent advancements in large scale contrastive learning has leveraged supervision from text to achieve high zero shot performance on down stream tasks [23, 15]. We evaluate the performance of the largest CLIP model[5] trained on 400 million image, text pairs from the internet. We evaluate two versions of this model, a "zero-shot" variant trained solely on 400 million images, text pairs and a "linear-probe" variant where the last linear layer was fine-tuned on ILSVRC-2012. We find that the zero shot variant while still suffering from a 6% accuracy drop is significantly more robust and accurate than any of the other models in our test bed. We note that due to the sheer amount of training data and the size of the model, these models are *incredibly* expensive to train and are out of reach to the computational resources of most researchers. Thus we leave further investigation of the robustness of these models to future work.

4.2. Detection

We further study the impact of natural perturbations on object detection. Specifically, we report results for two related tasks: object localization and detection. Object detection is the standard computer vision task of correctly classifying an object and finding the coordinates of a tight bounding box containing the object. "Object localization", meanwhile, refers to only the subtask of finding the bounding box, *without* attempting to correctly classify the object.

We provide our results on ImageNet-Vid-Robust, which contains dense bounding box labels unlike Youtube-BB, which only labels boxes at 1 frame per second. We use the popular Faster R-CNN [26] and R-FCN [5, 30] architectures for object detection and localization and report results in Table 3. For the R-FCN architecture, we use the model from [30][6]. We first note the significant drop in mAP of 12 to 15 points for object detection due to perturbed frames for both the Faster R-CNN and R-FCN architectures. Next, we show that localization is indeed easier than detection, as the mAP is higher for localization than for detection (e.g., 76.6 vs 62.8 for Faster R-CNN with a ResNet-50 backbone). Perhaps surprisingly, however, switching to the localization task does *not* improve the drop between original and perturbed frames, indicating that natural perturbations induce both classification and localization errors. We show examples of detection failures in Figure 4.

4.3. Impact of Dataset Review

We analyze the impact of our human review, described in Section 3.1, on the classifiers in our testbed. First, we compare the original and perturbed accuracies of a representative classifier (ResNet-152 finetuned) on frames with and without review in Section 4.1. We find that before review, the gap between the two accuracies is 16.2 and 10.0 on ImageNet-Vid-Robust and YTBB-Robust respectively. Our review improves the original accuracy by 3 to 4% (by discarding mislabeled or blurry anchor frames), and improves perturbed accuracy by 5 to 6% (by discarding dissimilar frame pairs). As a result, our review reduces the accuracy drop by 1.8% on ImageNet-Vid-Robust and 1.1% on YTBB-Robust. These results indicate that the changes in model predictions are indeed due to a lack of robustness, rather than due to significant differences between adjacent frames.

[5]The underlying model was a large visual transformer evaluated on 336 x 336 images (ViT-L/14@336px)

[6]This model was originally trained on the 2015 subset of ImageNet-Vid. We evaluated this model on the 2015 validation set because the method requires access to pre-computed bounding box proposals which are available only for the 2015 subset of ImageNet-Vid.

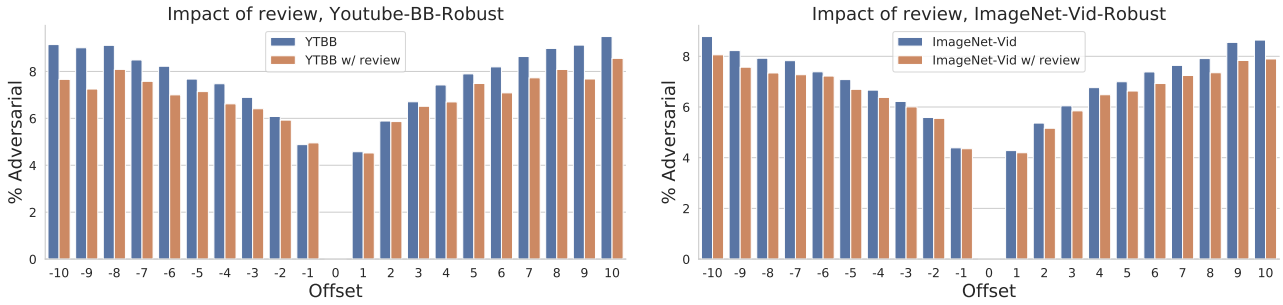

Figure 5: We plot how often each frame offset resulted in error, across all models, before and after review. Frames further away more frequently cause errors. Our review reduces errors by removing dissimilar frames, especially ones further away.

	Accuracy			
	Original	Perturbed	Δ	# anchors
All frames	84.8	70.2	14.6	1109
w/o 'i-frames'	84.7	70.3	14.4	1104
w/o 'p-frames'	83.9	73.7	10.2	415
w/o 'b-frames'	85.4	73.2	12.2	699

Table 5: Analyzing results based on compressed frame type (See Section 4.4).

To further analyze the impact of our review on model errors, we plot how frequently each offset distance from the anchor frame results in a model error across all model types in Figure 5. Larger offsets indicate pairs of frames further apart in time. For both datasets, we find that such larger offsets lead to more frequent model errors. Our review reduces the fraction of errors across offsets, especially for large offsets, which are more likely to display large changes from the anchor frame.

4.4. Video compression analysis

One concern with analyzing performance on video frames is the impact of video compression on model robustness. In particular, the 'mp4' videos in `ImageNet-Vid-Robust` contain 3 frame types: 'i-', 'p-', and 'b-' frames. 'p-frames' are compressed by referencing pixel content from previous frames, while 'b-frames' are compressed via references to previous and future frames. 'i-frames' are stored without references to other frames.

We compute the original and perturbed accuracies, as well as the accuracy drop for a subset without each frame type in Table 5. While there are modest differences in accuracy due to compression, our analysis suggests that the sensitivity of models is not significantly due to the differences in quality of frames due to video compression.

5. Conclusion

We analyze and quantify a common phenomenon in image models: flicker in predictions over time, which is caused by a lack of model robustness to natural perturbations. We show this results in significant accuracy drops for a wide range of classification and detection models. We highlight two key avenues for future research:

Building more robust models. Our benchmarks provide a standard robustness measure for classification and detection models. In Section 4.1, we found that several models suffer from substantial accuracy drops due to natural perturbations. Further, improvements with respect to artificial perturbations (like image corruptions or ℓ_∞ adversaries) induce only modest robustness improvements. One exception to this bleak overview are recent contrastive learning approaches trained on large-scale web data [23], which confer partial robustness to natural perturbations. We hope our standardized benchmarks will enable progress in improving the robustness of such models, and in generalizing their improvements to models trained on more limited datasets.

Further natural perturbations. Videos provide a straightforward method for collecting natural perturbations of images, enabling the study of realistic forms of robustness. Other methods for generating such natural perturbations are likely to provide additional insights into robustness. As an example, photo sharing websites contain many near-duplicate images: image pairs of the same scene captured at different times, viewpoints, or from a different camera [25]. More generally, devising similar, domain-specific strategies to collect, verify, and measure robustness to natural perturbations in domains such as natural language processing or speech recognition is a promising direction for future work.

Acknowledgements. We thank Rohan Taori for providing models trained for robustness to image corruptions, and Pavel Tokmakov for his help with training detection models on ImageNet-Vid. This research was generously supported in part by ONR awards N00014-17-1-2191, N00014-17-1-2401, and N00014-18-1-2833, the DARPA Assured Autonomy (FA8750-18-C-0101) and Lagrange (W911NF-16-1-0552) programs, an Amazon AWS AI Research Award, and a gift from Microsoft Research.

References

[1] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018. 2

[2] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 2018. https://arxiv.org/abs/1712.03141. 1, 2

[3] NTS Board. Collision between vehicle controlled by developmental automated driving system and pedestrian. *Nat. Transpot. Saf. Board, Washington, DC, USA, Tech. Rep. HAR-19-03*, 2019. 1

[4] Remi Cadene. Pretrained models for pytorch. https://github.com/Cadene/pretrained-models.pytorch. Accessed: 2019-05-20. 4

[5] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NeurIPS*, 2016. 7

[6] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017. 1, 2

[7] Alhussein Fawzi and Pascal Frossard. Manitest: Are classifiers really invariant? In *BMVC*, 2015. 1, 2

[8] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *ICCV*, 2017. 2

[9] 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. *ICLR*, 2019. 2

[10] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *NeurIPS*, pages 7538–7550, 2018. 2

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2

[12] Keren Gu, Brandon Yang, Jiquan Ngiam, Quoc Le, and Jonathan Shlens. Using videos to evaluate image model robustness. *arXiv preprint arXiv:1904.10076*, 2019. 2, 3, 6

[13] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1, 2, 4

[14] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *CVPR Workshop*, pages 1614–1619, 2018. 1, 2

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. 7

[16] SouYoung Jin, Aruni RoyChowdhury, Huaizu Jiang, Ashish Singh, Aditya Prasad, Deep Chakraborty, and Erik Learned-Miller. Unsupervised hard example mining from videos for improved object detection. In *ECCV*, 2018. 2

[17] Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Geometric robustness of deep networks: analysis and improvement. *arXiv preprint arXiv:1711.09115*, 2017. 2

[18] Kai Kang, Hongsheng Li, Tong Xiao, Wanli Ouyang, Junjie Yan, Xihui Liu, and Xiaogang Wang. Object detection in videos with tubelet proposal networks. In *CVPR*, 2017. 2

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*. Springer, 2014. 4

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018. 6

[21] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 3

[22] Harold Pashler. Familiarity and visual change detection. *Perception & psychophysics*, 44(4):369–378, 1988. 3

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2, 7, 8

[24] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *CVPR*, 2017. 2, 3

[25] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *ICML*, 2019. 2, 8

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 7

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015. 2, 3

[28] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *NeurIPS*, 2020. 5

[29] Antonio Torralba, Alexei A Efros, et al. Unbiased look at dataset bias. In *CVPR*, 2011. 2

[30] Fanyi Xiao and Yong Jae Lee. Video object detection with an aligned spatial-temporal memory. In *ECCV*, 2018. 2, 6, 7

[31] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. *arXiv preprint arXiv:1812.03411*, 2018. 5

[32] Haichao Zhang and Jianyu Wang. Towards adversarially robust object detection. In *ICCV*, pages 421–430, 2019. 2

[33] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *ICCV*, 2017. 2