Second Assignment

Literature Review in Computer Vision

Individual Contribution Declaration

Student Name	李骋昊	张如凡	杨昊运	周绍晅
Student ID	SA24011041	SA24011112	SA24011145	SA23011112
Search literature	✓			
Evaluate, Select Literature	✓	✓		
Identify Research Trend, Gaps, Debates, etc.		✓	✓	
Survey Structure			✓	✓
Writing				✓
Other Contributions (pls state):				

A Review of the Robustness of Deep Learning Models to Common Image Distortions and Corruptions

Chenghao Li
SA24011041

Rufan Zhang
SA24011112

Haoyun Yang
SA24011145

Shaoxuan Zhou
SA23011112

ABSTRACT

Deep learning models for image recognition, while powerful, are vulnerable to common image distortions (e.g., noise, blur). This review explores their robustness to such corruptions, covering definitions, benchmarks, and evaluation metrics. It surveys a wide range of enhancement techniques, from data augmentation and robust architectures to specialized training and adaptation strategies, while also investigating causes of model fragility. The review identifies critical research gaps and proposes future directions, including developing intrinsically robust models and more realistic benchmarks, to ensure reliable real-world deployment.

KEYWORDS

Deep Learning, Computer Vision, Robustness, Image Corruptions, Image Distortions, Benchmark

1 INTRODUCTION

In recent years, Deep Learning models, with Convolutional Neural Networks (CNNs)[42] and, more recently, Vision Transformers (ViTs)[11] at the forefront, have achieved state-of-the-art performance across a multitude of image recognition tasks. Their ability to learn hierarchical features directly from data has revolutionized fields ranging from object detection and image classification to semantic segmentation. However, the transition of these powerful models from controlled laboratory environments to real-world applications, such as autonomous driving, medical image analysis, industrial quality control, and security surveillance, is frequently challenged by their susceptibility to common image distortions and corruptions. These corruptions are not malicious, intentionally crafted attacks, but rather naturally occurring degradations that arise from various factors inherent in the image lifecycle: imperfections in camera sensors, suboptimal lighting conditions, motion blur, atmospheric phenomena, or artifacts introduced during image compression and transmission[47].

The practical implications of this lack of robustness are profound. In safety-critical systems like autonomous vehicles, a misinterpretation of a scene due to fog or sensor noise can lead to catastrophic accidents. Similarly, in medical diagnosis, where Deep Learning models are increasingly used to analyze scans, an inability to handle variations in image quality from different scanners or patient conditions could result in misdiagnosis, with severe consequences for patient outcomes[46]. The performance degradation is often substantial; models that exhibit high accuracy on pristine benchmark datasets can experience a dramatic drop in performance when confronted with even moderate levels of common corruptions[25]. For instance, studies have shown significant accuracy decreases and unreliable confidence estimations when models are evaluated on corrupted images[25], and even the most advanced architectures are not immune to these vulnerabilities[46].

1.1 Motivation and Scope

Despite substantial advancements in deep learning, achieving models that are consistently robust to common image corruptions remains a significant and persistent challenge[57]. Many contemporary models, celebrated for their high accuracy on standard, clean benchmark datasets, often exhibit a precipitous decline in performance when evaluated on inputs affected by these everyday degradations[46]. This discrepancy between performance on clean data and corrupted data highlights a critical gap in our ability to build truly reliable vision systems.

The motivation for this review stems from the urgent need to consolidate and critically assess the rapidly growing body of research dedicated to understanding and mitigating this robustness deficit. As deep learning models become increasingly integrated into critical aspects of society, their reliability under diverse real-world conditions is no longer a desirable attribute but a fundamental requirement. A comprehensive understanding of the current landscape—spanning the types of corruptions encountered, their quantifiable impact on model behavior, the benchmarks and metrics used for rigorous evaluation, and the diverse array of strategies developed to enhance robustness—is essential for guiding future research and informing best practices in model development and deployment. This rapid evolution necessitates periodic, up-to-date reviews to synthesize new knowledge and chart the path forward.

The scope of this literature review is to systematically survey and analyze the research pertaining to the robustness of deep learning models against common, naturally occurring image distortions and corruptions. Specifically, this review will cover:

- **Taxonomy of Common Image Corruptions:** A detailed classification and description of various types of image noise, blur, weather-related effects, and digital/algorithmic artifacts.
- **Impact on Model Performance:** An examination of how these corruptions affect the accuracy, confidence, and internal representations of deep learning models.
- **Standardized Benchmarks and Evaluation Metrics:** An overview of widely used datasets such as ImageNet-C, CIFAR-10-C, and ImageNet-P, along with metrics like mean Corruption Error (mCE), and a discussion of their strengths and limitations, including newer and more challenging benchmarks.
- **Performance of Deep Learning Architectures:** A comparative look at how different architectural paradigms, primarily CNNs and ViTs, perform under these corruptions.

- **Methods for Enhancing Corruption Robustness:** A comprehensive survey of data augmentation, model design, training strategies, and adaptation techniques that improve model robustness to image corruptions.

The review will strictly focus on common, naturally occurring corruptions and will not delve into adversarial attacks or defenses, except where techniques or insights from AR research have been explicitly adapted or found relevant to CR. The aim is to provide a holistic view of the problem, from characterization and evaluation to mitigation and fundamental understanding.

1.2 Organization of the Review

This literature review is structured to guide the reader from foundational concepts to advanced techniques and future outlooks concerning the robustness of deep learning models to common image distortions and corruptions.

Section 2 formally defines corruption robustness in the context of deep learning, elucidates the core challenge of performance degradation, and critically distinguishes common corruption robustness from the related but distinct field of adversarial robustness.

Section 3 forms the main body of this review and is divided into five principal subsections. It covers the definition and importance of corruption robustness, analyzes the typical effects of corruptions on model performance (including comparisons with human vision and frequency domain perspectives), presents a taxonomy of common image corruptions, provides an overview of standard benchmark datasets, and details the key evaluation metrics used in the field. Additionally, it offers a comprehensive survey of methodologies developed to improve model resilience. This section is further divided into Data Augmentation Strategies, Architectural Design Innovations, and Advanced Training Strategies, with detailed discussions of prominent techniques and their underlying principles within each category.

Section 4 synthesizes the findings from the literature survey. It presents critical insights, identifies significant research gaps and unsolved challenges, and outlines promising future research avenues that could lead to more robust and reliable deep learning systems.

Section 5 provides a summary of the key findings and their implications. It offers take-home messages for both researchers and practitioners interested in or working on the problem of common corruption robustness.

2 DESCRIPTION OF RESEARCH PROBLEM

In the domain of deep learning, corruption robustness refers to the ability of a model, typically a neural network, to maintain its performance (e.g., classification accuracy, object detection precision, segmentation quality) when its input images are subjected to common, often naturally occurring or incidental, distortions and corruptions[25]. These corruptions are not designed to be adversarial but rather represent the types of image degradation frequently encountered in real-world data acquisition and transmission processes. The term "statistical robustness" is sometimes used interchangeably to emphasize the non-adversarial, often stochastic nature of these perturbations[19].

Formally, let f be a deep learning model (e.g., a classifier) parameterized by θ, which maps an input image x from an input space $\mathcal{X}$ to an output y (e.g., a class label) in an output space $\mathcal{Y}$. Let $D_{\text{train}} = \{(x_i, y_i^*)\}$ be the training dataset consisting of clean images and their true labels. The model is typically trained to minimize a loss function $\mathcal{L}(f(x_i;\theta), y_i^*)$ over D_{train}.

A corruption can be represented by a transformation function $c : \mathcal{X} \to \mathcal{X}$, where $c(x)$ is the corrupted version of image x. Let C be a set of such corruption functions, representing various types and severities of common corruptions (e.g., Gaussian noise, motion blur, fog). Corruption robustness is then evaluated by assessing the performance of the trained model f on a test dataset D_{test} where each image $x \in D_{\text{test}}$ has been transformed by some $c \in C$. The core research problem arises from the empirical observation that the performance of f on the corrupted dataset $D_{\text{corr}} = \{(c(x), y^*) \mid (x, y^*) \in D_{\text{test}}, c \in C\}$ is often significantly worse than its performance on the clean D_{test}[69]. This performance degradation occurs even when the semantic content of $c(x)$ remains largely intact and easily discernible to human observers[16].

The challenge is thus to develop models f and training procedures such that the performance gap between clean and corrupted data is minimized. This involves understanding the nature of these corruptions, their impact on model representations and decision boundaries, and devising strategies to make models invariant or less sensitive to such input variations. The definition of "common" corruptions is often operationalized through standardized benchmark datasets like ImageNet-C[25], which includes a predefined set of 15 corruption types across 5 severity levels. While these benchmarks provide a valuable tool for standardized evaluation, they also implicitly shape the research focus. Consequently, models optimized for these specific benchmark corruptions may not necessarily exhibit robust behavior against novel or unmodeled types of common corruptions encountered in truly unconstrained environments, pointing to a persistent challenge in achieving broad out-of-distribution generalization.

2.1 Distinguishing Corruption Robustness from Adversarial Robustness

A critical distinction in the study of model robustness is between common corruption robustness and adversarial robustness[25]. While both involve evaluating model performance on perturbed inputs, the nature, origin, and intent of these perturbations differ significantly, leading to distinct research problems and defense strategies.

Common Corruptions:

- **Nature:** These are typically naturally occurring or incidental alterations to images. Examples include blur (motion, defocus), noise (Gaussian, shot), weather effects (fog, snow, rain), and digital artifacts (JPEG compression, pixelation)[49].
- **Origin:** They arise from imperfections in the image acquisition process (e.g., sensor noise, camera shake), environmental conditions, or data transmission and storage processes.
- **Intent:** Common corruptions are generally unintentional and not specifically designed to fool the model. They represent statistical deviations from the clean data distribution.

- **Visibility:** These corruptions are often visible and can range from subtle to severe, sometimes significantly degrading image quality but often leaving the core semantic content interpretable by humans.
- **Evaluation Focus:** The goal is to achieve good average-case performance across a diverse set of common corruption types and severities[25]. Robustness is measured by how well the model maintains its accuracy or other performance metrics on these generally perturbed inputs.
- **Model Specificity:** Common corruptions are typically model-agnostic; the same corruption (e.g., adding Gaussian noise with a certain standard deviation) can be applied to any image and evaluated against any model.

Adversarial Perturbations:

- **Nature:** These are small, often imperceptible, perturbations that are meticulously crafted and optimized to cause a specific target model to misclassify an input[41].
- **Origin:** They are intentionally generated by an adversary with knowledge (to varying degrees) of the target model.
- **Intent:** The intent is malicious ‒to induce a specific failure in the model, often while remaining undetectable to human observers.
- **Visibility:** Adversarial perturbations are typically designed to be quasi-imperceptible, falling within small (L_p)-norm bounds (e.g., (L_∞), (L_2)).
- **Evaluation Focus:** The goal is to achieve good worst-case performance against a specific attacker model and perturbation budget[19]. Robustness is measured by the model's ability to resist these optimized attacks.
- **Model Specificity:** Adversarial perturbations are often highly model-specific. An attack effective against one model architecture may not be as effective against another (though transferability exists).

Corruption robustness, in the context of deep learning, refers to a model's capacity to maintain its predictive performance when confronted with input images that have been degraded by common, non-adversarial distortions or corruptions[25]. These corruptions are statistical in nature and represent deviations from the idealized, clean data typically used for training. The importance of corruption robustness cannot be overstated, particularly as deep learning models are increasingly deployed in real-world applications where pristine data is the exception rather than the rule[52].

Unlike adversarial robustness, which is primarily a security concern focused on defending against meticulously crafted, worst-case attacks designed to deceive a model[25], corruption robustness addresses a more pervasive and often unavoidable challenge: the natural variability and degradation of image quality due to environmental factors, sensor limitations, and processing artifacts. For instance, autonomous vehicles must operate reliably in diverse weather conditions like rain, fog, or snow, and contend with sensor noise or motion blur[45]. In medical imaging, diagnostic models need to be resilient to artifacts arising from image acquisition, patient movement, or digitization errors[41]. Failure to maintain performance under such conditions can lead to incorrect decisions, reduced efficacy, and, in safety-critical systems, potentially catastrophic outcomes.

The research community largely treats these as related but distinct subfields[57]. Strategies that enhance adversarial robustness, such as adversarial training, do not always translate to improved robustness against common corruptions, and vice-versa[19]. For instance, adversarial training might make a model overly sensitive to the specific patterns of adversarial noise, potentially harming its generalization to the broader, more stochastic patterns of common corruptions.

Conversely, defenses focused on common corruptions by, for example, smoothing high-frequency details might not be sufficient against finely tuned adversarial attacks. This review specifically concentrates on the challenges and solutions related to common image corruptions, a problem of immense practical importance for the deployment of reliable AI systems in uncontrolled, real-world settings where data quality cannot be guaranteed. The prevalence of such corruptions in critical applications, like medical imaging[69] and autonomous driving[45], underscores the urgency of this research area, as performance degradation can directly impact safety and efficacy. The pursuit of robustness against common corruptions is fundamentally a quest for models that generalize better to the inherent variability of the visual world, a goal that remains elusive despite considerable research efforts.

3 LITERATURE SURVEY

3.1 Taxonomy of Common Image Corruptions

To systematically study and address the impact of common image corruptions, researchers have developed taxonomies to categorize them. A widely adopted categorization was introduced by Hendrycks and Dietterich in their work on the ImageNet-C benchmark[25]. They grouped 15 algorithmically generated common corruptions into four primary types, each with multiple severity levels:

- **Noise Corruptions:** These involve adding random pixel-level variations.
  - Gaussian Noise: Additive noise drawn from a Gaussian distribution.
  - Shot Noise (Poisson Noise): Noise that follows a Poisson distribution, often modeling sensor noise in low-light conditions.
  - Impulse Noise (Salt-and-Pepper Noise): Random occurrences of black and white pixels.
- **Blur Corruptions:** These simulate various forms of image unfocusing or smearing.
  - Defocus Blur: Simulates an out-of-focus camera lens.
  - Frosted Glass Blur: Simulates viewing an image through frosted glass.
  - Motion Blur: Simulates blur due to camera or object movement.
  - Zoom Blur: Simulates blur caused by zooming during exposure.
- **Weather Corruptions:** These mimic degradations caused by atmospheric conditions.
  - Snow: Simulates snowfall, adding snowflakes and reducing visibility.
  - Frost: Simulates ice crystals forming on a surface, obscuring details.

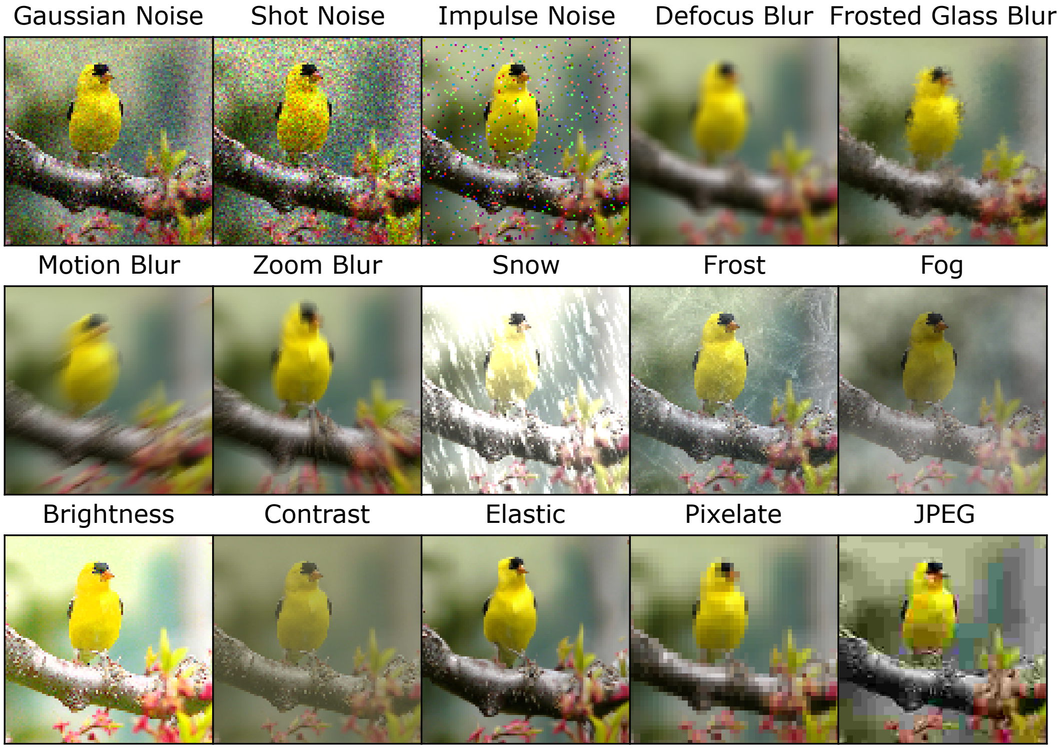

Figure 1: IMAGENET-C dataset consists of 15 types of algorithmically generated corruptions from noise, blur, weather, and digital categories. Each type of corruption has five levels of severity, resulting in 75 distinct corruptions.

  - Fog: Simulates reduced visibility due to fog, decreasing contrast and adding a haze. (Brightness is sometimes included here or as a digital corruption).
- **Digital Corruptions:** These represent artifacts introduced during digital image processing or transmission.
  - Brightness: Uniformly increases or decreases image brightness.
  - Contrast: Increases or decreases the difference between light and dark areas.
  - Elastic Transformation: Applies local elastic deformations to the image.
  - Pixelate: Reduces image resolution by averaging pixel values within blocks.
  - JPEG Compression: Simulates artifacts introduced by JPEG lossy compression.

Beyond this general taxonomy, domain-specific corruptions are also critical in certain applications. For example, in the field of digital pathology, Zhang et al.[69] identified relevant corruptions including:

- Digitization Artifacts: Such as those from JPEG compression and pixelation specific to scanned slides.
- Blur Types: Defocus and motion blur relevant to microscope imaging.
- Color Variations: Changes in brightness, saturation, and hue due to staining or scanning inconsistencies.
- Stain-Specific Artifacts: Such as marks or bubbles on the slide.

Other taxonomies might also consider geometric transformations (e.g., rotation, scaling, shearing) and various sensor-specific errors or structured noise patterns, such as those simulating faulty sensor lines[41]. While the ImageNet-C taxonomy has become a de facto standard for benchmarking general-purpose vision models, the consideration of domain-specific corruptions is crucial for developing robust systems tailored to particular real-world tasks.

Table 1 summarizes the common image corruptions based primarily on the ImageNet-C framework and digital pathology benchmarks, as these are extensively referenced in the literature on DL robustness.

3.2 Benchmark Datasets

The development and use of standardized benchmark datasets have been instrumental in advancing research on common corruption robustness. These benchmarks provide a common ground for evaluating model performance, comparing different robustness-enhancing techniques, and tracking progress in the field.

Foundational Benchmarks:

- **ImageNet-C and ImageNet-P:** Introduced by Hendrycks and Dietterich[25], ImageNet-C is arguably the most influential benchmark for common corruption robustness in image classification. It consists of the ImageNet validation

Table 1: Taxonomy of Common Image Corruptions

Category	Specific Corruption Type	Brief Description
Noise	Gaussian Noise	Additive noise from a Gaussian distribution affecting pixel intensity.
	Shot Noise	Noise following a Poisson distribution, common in low-light imaging.
	Impulse Noise	Random white and black pixels (salt-and-pepper).
Blur	Defocus Blur	Simulates an out-of-focus camera lens.
	Frosted Glass Blur	Simulates viewing an image through a textured glass surface.
	Motion Blur	Simulates blur due to relative motion between camera and subject.
	Zoom Blur	Simulates blur caused by changing focal length during exposure.
Weather	Snow	Adds simulated snowflakes and reduces visibility.
	Frost	Simulates ice crystal formation on a surface, obscuring details.
	Fog	Reduces contrast and adds a haze effect, simulating foggy conditions.
Digital	Brightness	Uniformly alters the overall lightness or darkness of the image.
	Contrast	Adjusts the difference between light and dark areas.
	Elastic Transformation	Applies local, non-linear deformations to the image structure.
	Pixelate	Reduces image resolution by averaging pixel values within blocks.
	JPEG Compression	Introduces artifacts typical of lossy JPEG compression.
Pathology-Specific	Saturation	Adjusts the intensity of colors in pathology slides.
	Hue	Shifts the overall color palette, relevant for stain variations.
	Mark	Simulates pen marks or other annotations on slides.
	Bubble	Simulates air bubbles trapped under the coverslip of a slide.

set corrupted by 15 diverse, algorithmically generated corruption types (categorized into noise, blur, weather, and digital), each applied at five distinct severity levels. This results in 75 unique corruption variations. ImageNet-P, released concurrently, focuses on perturbation robustness, evaluating model consistency on sequences of images with small, evolving perturbations. These datasets standardized the evaluation methodology and revealed the brittleness of then-current models.

- **CIFAR-10-C and CIFAR-100-C:** Following the methodology of ImageNet-C, corrupted versions of the CIFAR-10 and CIFAR-100 datasets were created. These provide smaller-scale benchmarks for faster experimentation and are widely used for evaluating models trained on these respective datasets.

Domain-Specific Benchmarks: The need for robustness in specific application domains has led to the creation of tailored benchmarks:

- **Digital Pathology:** Zhang et al.[69] developed Patchcamelyon-C and LocalTCT-C by applying a set of common and pathology-specific corruptions (e.g., stain variations, marks, bubbles) to the validation sets of the Patchcamelyon and LocalTCT datasets, respectively. These benchmarks are crucial for evaluating diagnostic models intended for clinical use.
- **Person Search:** To address robustness in person re-identification and detection under real-world surveillance conditions, benchmarks like CUHK-SYSU-C and PRW-C have been proposed, incorporating corruptions relevant to this task[51].
- **Autonomous Driving:** Beyond Cityscapes-C, researchers utilize datasets that capture real-world adverse conditions, including varied weather and sensor noise, to evaluate perception systems for autonomous vehicles[49].

Newer and More Diverse Benchmarks: The field continues to evolve, with newer benchmarks aiming to address limitations of earlier ones or to test different facets of robustness:

- **PASCAL VOC-C, COCO-C, and Cityscapes-C:** Recognizing that robustness is critical beyond classification, Michaelis et al.[38] extended the ImageNet-C corruption types to popular object detection and semantic segmentation datasets, namely PASCAL VOC[14], MS COCO[33], and Cityscapes[6]. These benchmarks (PASCAL-C, COCO-C, Cityscapes-C) allow for the assessment of corruption robustness in more complex vision tasks, which is particularly relevant for applications like autonomous driving.
- **ImageNet-C̄ and CIFAR-C̄:** These datasets, mentioned in the survey by Mintun et al.[39], were created with corruptions designed to be perceptually dissimilar to those in the original ImageNet-C and CIFAR-10-C. The goal is to test for broader generalization to unseen corruption types rather than just performance on a fixed set.
- **ImageNet-R:** Introduced by Hendrycks et al.[24], ImageNet-R consists of artistic renditions (e.g., paintings, cartoons, sculptures, sketches, tattoos) of ImageNet object classes. It tests robustness to significant variations in texture, style, and local image statistics, moving beyond typical noise and blur corruptions.
- **ImageNet-D:** Proposed by Zhang et al.[65], ImageNet-D leverages advanced diffusion models to generate images

with highly diversified backgrounds, textures, and materials for ImageNet classes. It presents a new level of challenge by creating high-fidelity synthetic images that can significantly degrade the performance of even robust models.

- **ImageNet-3DCC:** This benchmark[29] introduces corruptions that consider the 3D geometry of scenes, offering a more realistic simulation of certain real-world distortions compared to purely 2D image-space corruptions.

3.3 Evaluation Metrics

Quantifying the robustness requires well-defined evaluation metrics. These metrics aim to capture different aspects of model performance under duress, from simple accuracy degradation to more nuanced measures of reliability and consistency.

3.3.1 Error-Based Metrics.

- **Corruption Error (CE):** This is the most straightforward metric, representing the classification error rate (or task-specific error) of a model on a dataset corrupted by a specific type and severity of distortion. It is often reported for each individual corruption.
- **Mean Corruption Error (mCE):** Popularized by the ImageNet-C benchmark[25], mCE is the most common metric. It normalizes a model's error rate on each corruption type and severity by a baseline model's (typically AlexNet) error, then averages these normalized errors. Let $E_{f,s,c}$ be the error of model f on corruption c at severity s. The Corruption Error for corruption c is $\text{CE}_{f,c} = \left(\sum_{s=1}^{N_s} E_{f,s,c}\right) / \left(\sum_{s=1}^{N_s} E_{\text{AlexNet},s,c}\right)$. Then, $\text{mCE}_f = \frac{1}{N_c} \sum_{c=1}^{N_c} \text{CE}_{f,c}$. Lower mCE is better.
- **Relative Mean Corruption Error (rCE):** Also from Hendrycks and Dietterich[25], this measures degradation relative to clean data performance, again normalized by AlexNet's relative degradation. Relative $\text{mCE}_f = \frac{1}{N_c} \sum_{c=1}^{N_c} \frac{\sum_{s=1}^{N_s} (E_{f,s,c} - E_{f,\text{clean}})}{\sum_{s=1}^{N_s} (E_{\text{AlexNet},s,c} - E_{\text{AlexNet},\text{clean}})}$. This revealed that mCE gains in newer models were often due to higher clean accuracy, not better resilience[25].

3.3.2 Perception and Dataset-Aware Metrics.

- **Visually-Continuous Corruption Robustness (VCR) Metrics:** Proposed by Shen et al.[48], VCR aims to assess robustness across a continuous spectrum of visual degradation, rather than at fixed severity levels. It uses an Image Quality Assessment (IQA) metric, specifically Visual Information Fidelity (VIF), to quantify the perceptual change (Δv) caused by a corruption. Key VCR metrics include:
  - Accuracy (R_a): The expected probability that the model's prediction on corrupted images is correct, integrated over the continuous range of visual corruption (Δv).
  - Prediction Consistency (R_p): The expected probability that the model's prediction on a corrupted image is the same as its prediction on the original (clean) image, integrated over the continuous range of (Δv).

  To compare neural network VCR with human performance, two novel indices are introduced: Human-Relative Model Robustness Index (HMRI), which quantifies how well an NN replicates human VCR, and Model Robustness Superiority Index (MRSI), which measures the extent to which an NN exceeds human VCR. A limitation noted is that VCR, using VIF, is primarily suited for pixel-level corruptions[48].
- **Mean Statistical Corruption Robustness (MSCR):** Introduced by Jung et al.[19], MSCR is a dataset-specific metric designed for comparability and interpretability. It first determines a ”robustness distance” (ϵ_{min}) derived from the minimal class separation distance within the dataset. Test data is then augmented by adding noise within this (ϵ_{min}) radius. The MSCR is calculated as the normalized difference between the robust accuracy on this augmented data ($\text{Acc}_{\text{rob}-\epsilon_{\text{min}}}$) and the clean accuracy ($\text{Acc}_{\text{clean}}$): ($MSCR = (Acc_{\text{rob}-\epsilon_{\text{min}}} - Acc_{\text{clean}})/Acc_{\text{clean}}$). A higher MSCR indicates better corruption robustness relative to the dataset's inherent separability.

3.3.3 Confidence and Consistency Metrics.

- **Expected Calibration Error (ECE):** As corruptions can affect not only accuracy but also the reliability of a model's confidence scores[69], ECE is an important metric[12]. It measures the discrepancy between a model's predicted confidence and its actual accuracy, indicating how well-calibrated the model's probabilities are.
- **Corruption Error of Confidence (CEC):** Designed for digital pathology benchmarks by Zhang et al.[69], CEC measures the mismatch between predictive confidence values and the intuition that confidence should generally decrease with increasing corruption severity.
- **Mean Flip Rate (mFR) and Mean Top-5 Distance (mT5D):** These metrics are particularly relevant for perturbation benchmarks like ImageNet-P[12]. mFR measures how often the top-1 prediction changes between consecutive frames in a perturbation sequence. mT5D measures the average change in the set of top-5 predictions, indicating stability.

The choice of metric depends on the specific aspect of robustness being investigated. While mCE has been a standard for ImageNet-C, newer metrics like VCR and MSCR attempt to provide more nuanced or context-aware evaluations. The increasing attention to calibration (ECE) and confidence (CEC) reflects a growing demand for models that are not only accurate but also reliable in their self-assessment, especially under challenging input conditions.

The evolution of these metrics from simple error rates to more sophisticated measures that incorporate perceptual quality, dataset characteristics, and confidence reliability signifies a maturing field. This progression is crucial for developing a deeper understanding of model behavior and for guiding the creation of truly robust AI systems that can reliably operate in the complexities of the real world. The persistent gap in performance between models on clean versus corrupted data, and the often significant disparity between AI and human robustness[48], underscores that this is a deep-seated issue related to how current deep learning models learn and generalize from data. The problem is fundamentally one of out-of-distribution (OOD) generalization[52], where corruptions represent specific types of distribution shifts that violate the

Table 2: Overview of Benchmark Datasets for Corruption Robustness

Dataset Name	Base Dataset(s)	Year Intro.	Characteristics / Corruption Types	Reference(s)
ImageNet-C	ImageNet	2019	15 common synthetic corruptions at 5 severity levels.	[10, 25]
CIFAR-10-C	CIFAR-10	2019	Same 15 corruptions and 5 severities as ImageNet-C.	[25, 31]
CIFAR-100-C	CIFAR-100	2019	Same 15 corruptions and 5 severities as ImageNet-C.	[25, 31]
ImageNet-P	ImageNet	2019	10 common perturbation types applied as sequences (video-like).	[10, 25]
PASCAL-C	PASCAL VOC	2020	ImageNet-C corruption methodology applied to PASCAL VOC.	[14, 38]
COCO-C	MS COCO	2020	ImageNet-C corruption methodology applied to MS COCO.	[33, 38]
Cityscapes-C	Cityscapes	2020	ImageNet-C corruption methodology applied to Cityscapes.	[6, 38]
ImageNet-C̄	ImageNet	2021	10 new corruptions perceptually dissimilar to ImageNet-C.	[10, 39]
ImageNet-R	ImageNet	2021	Renditions of ImageNet classes (e.g., art, cartoons, paintings, sculptures).	[10, 24]
ImageNet-3DCC	ImageNet	2022	Realistic synthetic corruptions using 3D geometry.	[10, 29]
Patchcamelyon-C	Patchcamelyon	2022	ImageNet-C corruption methodology applied to Patchcamelyon.	[56, 69]
CUHK-SYSU-C	CUHK-SYSU	2024	Synthetic corruptions (ImageNet-C style) for person re-identification.	[51, 60]
PRW-C	PRW	2024	Synthetic corruptions (ImageNet-C style) for person re-identification.	[51, 71]
ImageNet-D	ImageNet	2024	Images of ImageNet classes from diverse real-world domains.	[10, 65]
TextVQA-C	TextVQA	2025	ImageNet-C style corruptions applied to images for VQA.	[50, 54]
GQA-C	GQA	2025	ImageNet-C style corruptions applied to images for VQA.	[27, 54]

Table 3: Key Evaluation Metrics for Corruption Robustness

Metric Name	Definition/Formula	Purpose/Interpretation
Corruption Error (CE)	Error rate on a specific corruption type/severity.	Measures raw performance on corrupted data.
Mean Corruption Error (mCE)	Average of error rates on corrupted data, normalized by a baseline model's error, across types/severities.	Standardized measure of relative robustness; lower is better.
Relative Corruption Error (rCE)	Performance drop on corrupted vs. clean data, often normalized.	Quantifies relative degradation for a specific model.
VCR - Accuracy (R_a)	Expected probability of correct prediction over continuous visual corruption range (via IQA).	Measures accuracy robust to continuous perceptual changes.
VCR - Consistency (R_p)	Expected probability of consistent prediction (corrupted vs. original) over continuous visual corruption range.	Measures prediction stability under continuous perceptual changes.
Mean Statistical Corruption Robustness (MSCR)	$((Acc_{\text{rob}-\epsilon_{\min}} - Acc_{\text{clean}})/Acc_{\text{clean}})$, where $(\epsilon_{\min})$ is from class separation.	Dataset-aware metric; higher indicates better robustness relative to dataset difficulty.
Expected Calibration Error (ECE)	Weighted average of difference between accuracy and confidence in predefined bins.	Measures reliability of confidence scores; lower is better calibrated.
Corruption Error of Confidence (CEC)	Measures mismatch between confidence and expectation that confidence decreases with corruption severity.	Assesses reliability of confidence trends under increasing corruption.
Mean Flip Rate (mFR)	Average frequency of top-1 prediction changes in perturbation sequences.	Measures prediction stability against small, evolving perturbations.

i.i.d. assumption made during training. Models tend to learn spurious correlations present in the clean training data, which do not hold when the data is corrupted, leading to failures[52].

3.4 Impact of Corruptions on Model Performance

Generally, deep models are surprisingly brittle to common degradations, with error rates often doubling or tripling[25]. Different architectures show varying robustness, but even modern ViTs are not immune[21]. Vulnerabilities are corruption-specific, highlighting complex model failure modes. For VLMs, text recognition is most affected by blur/snow, while object reasoning tasks are more sensitive to frost/impulse noise, linked to frequency characteristics and Transformer biases[55].

Performance Degradation: Standard deep learning models, including those considered state-of-the-art on clean benchmark datasets, exhibit a significant drop in accuracy and other performance metrics when evaluated on corrupted images[69]. Michaelis et al.[37] reported that the performance of object detection models can plummet to as low as 30-60% of their original performance on clean data when faced with common corruptions. Similarly, Zhang et al.[69] found that Deep Neural Networks (DNNs) used for digital pathology images experienced a substantial decrease in accuracy

—effectively doubling their error rate compared to clean images— and exhibited unreliable confidence estimations when inputs were corrupted. This degradation is not limited to older architectures; even modern networks show this vulnerability.

Model Confidence Issues: Beyond merely reducing accuracy, image corruptions can also adversely affect the reliability of a model's confidence scores. Ideally, a model's confidence should decrease when presented with ambiguous or severely degraded input. However, studies have shown that models can become overconfident in their incorrect predictions on corrupted data[69]. In some counter-intuitive instances, model confidence has been observed to increase with the severity of certain corruptions, further undermining the trustworthiness of their outputs[69].

Comparison with Human Visual Robustness: A striking aspect of this problem is the disparity between the robustness of deep learning models and that of the human visual system. Humans possess a remarkable ability to perceive and interpret scenes accurately despite various forms of image degradation, such as blur, noise, or adverse weather[48]. Current deep learning models, in contrast, struggle significantly more with these same corruptions[48]. Shen et al.[48], through their proposed Visually-Continuous Corruption Robustness (VCR) framework, quantified this gap and concluded that it is even larger than previously understood. Their findings indicate that no existing neural network fully matches human performance across the continuous spectrum of corruption levels, either in terms of accuracy or prediction consistency. Furthermore, the types of errors made by humans and neural networks on corrupted images are often uncorrelated, suggesting fundamental differences in their underlying processing mechanisms and reliance on different visual cues[48].

Frequency Domain Analysis: The impact of corruptions can also be analyzed from a frequency domain perspective. Certain corruptions predominantly affect specific frequency components of an image. For example, research has shown that fog and contrast alterations tend to impact low-frequency components, while noise-related corruptions (e.g., Gaussian noise, shot noise) primarily introduce high-frequency artifacts[51]. Corruptions like blur and pixelation typically affect mid-frequency components. Models that exhibit a strong bias towards high-frequency details, often referred to as "texture bias," may be particularly vulnerable to corruptions that disrupt these textures or, conversely, may fail when high-frequency information is less reliable[23]. This suggests that the way models process and prioritize information across different spatial frequencies plays a crucial role in their robustness to various types of common corruptions. Understanding these frequency-specific effects can inform the design of more robust architectures and training strategies.

3.5 Enhancing Robustness to Common Corruptions

Addressing the vulnerability of deep learning models to common image corruptions has been a major focus of research. A wide array of techniques has been proposed, broadly falling into three categories: data augmentation strategies, architectural design innovations, and advanced training strategies. These approaches aim to either expose the model to more diverse data, build inherent resilience into the model's structure, or guide the learning process towards more robust feature representations. The comprehensive survey by Wang et al.[57] provides an excellent taxonomy and detailed review of many of these methods, forming a basis for the organization of this section.

3.5.1 Data Augmentation Strategies. Data augmentation is arguably the most extensively explored and often most effective category of methods for improving robustness to common corruptions[12]. The fundamental rationale is that by exposing models to a wider and more diverse range of data variations during training, they are encouraged to learn features that are invariant to superficial changes, thereby generalizing better to unseen corrupted data encountered during testing[45].

Basic and Mixing-based Augmentations: Traditional data augmentation techniques include simple geometric and color transformations such as flipping, random cropping, rotation, translation, and color jittering. Kernel-based filters can be used to simulate blur, while pixel erasing or adding random noise (e.g., Gaussian, speckle) directly introduces forms of corruption into the training data[57]. Patch Gaussian augmentation applies Gaussian noise to small, randomly selected image patches rather than the entire image, which has been shown to benefit classification on both clean and corrupted images[57].

More advanced are mixing-based augmentations:

- **Mixup:** Proposed by Zhang et al.[66], Mixup creates new training samples by taking convex combinations of pairs of images and their corresponding labels: $\tilde{x} = \lambda x_i + (1-\lambda)x_j$ and $\tilde{y} = \lambda y_i + (1-\lambda)y_j$, where $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$. This encourages the model to behave linearly between training samples, improving generalization and offering some robustness, particularly to label noise[12].
- **CutMix:** Introduced by Yun et al.[64], CutMix involves cutting a patch from one image and pasting it onto another, with labels mixed proportionally to the area of the patches. This forces the model to learn from non-local information and to recognize objects even when parts are occluded or replaced, thereby enhancing localization ability and robustness[12]. Other mixing strategies like SmoothMix, GuidedMixup, and PuzzleMix further refine how images are combined[12]. LISA (Learning Invariant and Scrambled Augmentations) focuses on intra-label and intra-domain augmentations[12].

Learned Augmentation Policies: Recognizing that manually designing optimal augmentation strategies is challenging, methods have been developed to learn them:

- **AutoAugment:** Cubuk et al.[7] used reinforcement learning to search for an optimal policy consisting of sequences of augmentation operations (e.g., rotation, shear, color adjustments) and their magnitudes that maximize validation accuracy on a target dataset. While primarily aimed at improving clean accuracy, AutoAugment policies have also been shown to improve corruption robustness[12]. Subsequent work like Adversarial AutoAugment, PRIME (Policy

Table 4: Overview of Major Strategies for Enhancing Corruption Robustness

Category	Prominent Techniques	Reference(s)
Data Augmentation	Basic Transforms	
	Mixup	[66]
	CutMix	[64]
	AutoAugment	[7]
	AugMix	[26]
	Stylized-ImageNet	[18]
	DAMP	[53]
	Generative Models (GANs, Diffusion)	
	Test-Time Augmentation/Adaptation	
Architectural Design	BlurPool	[67]
	Adaptive Normalization (AdaBN, PAN)	[2]
	Robust Vision Transformer (RVT)	[36]
Training Strategies	Adversarial Logit Pairing (ALP)	[28]
	Sharpness-Aware Minimization (SAM)	[17]
	Contrastive Learning (SimCLR, GPaCo)	[5, 8]
	Knowledge Distillation (NoisyStudent, DAD)	[62, 72]

Regularized Implicit Maximum Entropy)[40], AugMax (adversarially selecting augmentations), and ME-AdA (Maximum-Entropy-based Adversarial Augmentation) have further explored learning or adversarially finding effective augmentation policies[12].

Advanced Compositional and Adversarial Augmentations: These methods often involve more complex transformations or explicitly aim for robustness:

- **AugMix:** Developed by Hendrycks et al.[26], AugMix stochastically samples and layers simple augmentation operations (e.g., translation, shear, solarize, posterize, equalize) into multiple "augmentation chains." The resulting diverse augmented images are then mixed with each other and with the original image using convex combinations. A crucial component is a Jensen-Shannon Divergence (JSD) consistency loss, which enforces that the model produces similar predictions for different augmented versions of the same input image. AugMix has demonstrated significant improvements in mean Corruption Error (mCE) on benchmarks like ImageNet-C and CIFAR-10-C, often without degrading clean accuracy[26]. It is widely referenced as a strong baseline for corruption robustness[12].
- **DeepAugment:** Also from Hendrycks et al.[48], DeepAugment applies augmentations in the deep feature space of an image. It uses image-to-image neural networks (e.g., autoencoders) and perturbs their weights or activations during a forward pass on a clean image. This process generates a wide variety of semantically consistent yet structurally diverse distortions. DeepAugment has shown strong performance on ImageNet-C and particularly on ImageNet-R (which tests robustness to artistic renditions)[24]. It is often combined with AugMix (referred to as DA+AM or AMDA) to achieve state-of-the-art robustness results on common corruption benchmarks[24].
- **Stylized-ImageNet:** Geirhos et al.[18] proposed training models on Stylized-ImageNet, a version of ImageNet where original image textures are replaced by various artistic styles using AdaIN style transfer. This encourages models to develop a "shape bias," relying more on global object shapes rather than superficial texture cues, which has been shown to improve robustness to certain corruptions and out-of-distribution generalization[22].
- **Data Augmentation via Multiplicative Perturbations (DAMP):** Trinh et al.[53] introduced DAMP, based on the premise that input perturbations can be mimicked by applying multiplicative perturbations to the model's weights during training. Random Gaussian noise is multiplied with the weights, aiming to improve corruption robustness without significantly impacting clean image accuracy or increasing computational overhead. DAMP has shown promise, especially for training Vision Transformers from scratch[53].
- **AdversarialAugment:** This technique[1] takes an adversarial approach by optimizing the parameters of image-to-image models (like autoencoders) to generate "adversarially corrupted" images. These images are then used for data augmentation, with the goal of improving robustness to both common corruptions and, to some extent, adversarial perturbations.

Frequency Domain Augmentations: Given that many corruptions have distinct frequency characteristics, manipulating images in the frequency domain has emerged as an augmentation strategy. This can involve mixing amplitude and phase spectra of different images or applying specific filters. Examples cited in Razzaq et al.[12] include APR-SP (Amplitude-Phase Recombination for Self-Supervision), VIPAug (Vital Phase Augmentation), AFA (Adversarial Frequency Augmentation), HybridAugment++, AugSVF (AugMix with Fourier noise), and RobustMix. Research by Chan et al.[4] explored how directly biasing classifiers towards specific frequency components (via Jacobian regularization rather than direct data augmentation) impacts robustness, finding that a preference for low spatial frequencies can improve robustness to high-frequency corruptions, and vice-versa[23].

Generative Models (GANs, Diffusion Models) for Data Augmentation: Generative models offer powerful tools for synthesizing diverse training data:

- Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have been used to mimic corruption effects or perform style transfer to generate augmented data.
- More recently, Diffusion Models have demonstrated exceptional capabilities in generating high-fidelity and highly diverse synthetic images. They are being explored for creating images with specific corruptions or variations that can challenge and improve model robustness[12]. The ImageNet-D benchmark, for instance, uses diffusion models to create challenging test images with varied backgrounds, textures, and materials[65].

Test-Time Augmentation (TTA) and Test-Time Adaptation (TTAda): Instead of (or in addition to) augmenting training data, transformations can be applied at test time:

- **Test-Time Augmentation (TTA):** This involves creating multiple augmented versions of a single test image, obtaining predictions for each, and then aggregating these predictions (e.g., by averaging) to yield a final, potentially more robust output[12]. TTA can improve both accuracy and calibration.
- **Test-Time Adaptation (TTAda):** This more dynamic approach involves adapting the model's parameters (e.g., Batch Normalization statistics, or even parts of the full model) at test time using the incoming unlabeled test data. Common objectives for adaptation include minimizing prediction entropy (TENT)[63], using pseudo-labels, or ensuring consistency across augmentations[12].
- **Test-time Enhancer and Classifier Adaptation (TECA):** Proposed by Fujihara et al.[13], TECA uniquely combines an image enhancement model with the classification model. Both models are updated at test time, with a "logit switching" mechanism selecting the prediction (from original or enhanced image) with lower uncertainty. TECA has demonstrated improved accuracy on ImageNet-C by effectively reducing prediction uncertainty.

The evolution from simple handcrafted augmentations to sophisticated learned policies and generative approaches highlights a continuous effort to bridge the gap between clean training distributions and the diverse, corrupted data encountered in the real world. The most successful strategies often create a rich and varied training experience for the model, forcing it to learn more fundamental and invariant features.

3.5.2 Architectural Design Innovations. While data augmentation focuses on diversifying the input data, another avenue for enhancing robustness to common corruptions lies in modifying the neural network architecture itself. The premise is that certain architectural choices can make models inherently more resilient to input variations and less prone to relying on spurious or fragile features.

Robust Convolutional Components: Within CNNs, specific components have been identified or redesigned for better robustness:

- **BlurPool:** Proposed by Zhang[67], BlurPool addresses the issue of shift-variance in traditional CNNs caused by max-pooling and strided convolutions. It replaces these operations with a low-pass blur filter (e.g., binomial, triangle kernel) followed by a downsampling step (e.g., strided convolution with stride > 1 or average pooling). By explicitly blurring before downsampling, BlurPool helps to satisfy the Nyquist sampling theorem more closely, reducing aliasing and improving the network's shift-invariance (or, more accurately, shift-equivariance). This increased stability to small input shifts can translate to better robustness against certain types of corruptions that might manifest as local translations or high-frequency artifacts. The survey by Razzaq et al.[12] reports a mean Corruption Error (mCE) of 58.1% on ImageNet-C for a ResNet-50 equipped with BlurPool.
- **Other Kernel and Receptive Field Modifications:** Research has explored altering the receptive fields of convolutional kernels or using biologically-inspired kernel designs. Examples include the Push-pull Layer, which uses kernels inspired by simple and complex cells in the visual cortex; on-off kernels designed for robust responses; and smoothing kernels aimed at stabilizing feature maps[12]. These approaches attempt to build in robustness at a more fundamental level of feature extraction.

Normalization Layers: Normalization layers, particularly Batch Normalization (BN), play a crucial role in training deep networks, but their statistics (running mean and variance) are typically estimated from the clean training data. When test data comes from a different distribution (e.g., due to corruptions), these stored statistics can be suboptimal.

- **Adaptive Batch Normalization (AdaBN) / Test-Time BN Adaptation:** One strategy is to adapt the BN statistics at test time using the statistics of the incoming test batch[12, 67].This allows the normalization to better match the current data distribution, which can improve performance under distribution shifts, including some common corruptions. CorrectBN is a variant that attempts to correct BN statistics using information from both training and test data[12].
- **Per-corruption Adaptation of Normalization (PAN):** The PAN framework[2] takes this further by proposing dynamic adjustment of normalization layer statistics based on an identified corruption type in the input image. It involves a corruption identification module and a mechanism to update normalization statistics in real-time according to the input. This approach recognizes that different corruptions might require different normalization adjustments. The core observation is that BN statistics differ significantly

for images with different corruption types but are similar for images with the same corruption type[2].

Robust Transformer Architectures: Vision Transformers (ViTs) have emerged as powerful alternatives to CNNs, offering different inductive biases. Their robustness characteristics are an active area of research. While standard ViTs might be inherently more robust to certain texture changes due to their global attention mechanisms, they can be vulnerable to perturbations affecting input patches or positional encodings.

- **Robust Vision Transformer (RVT):** Proposed by Mao et al.[36], RVT aims to enhance the robustness of ViTs. The design is based on systematically evaluating ViT components and identifying those detrimental to robustness (e.g., the standard classification token and locality constraints in patch attention). RVT incorporates ”robust building blocks,” such as modifications to patch embedding and position-aware attention scaling. It also employs specific training strategies like patch-wise augmentation. The goal is to achieve superior performance with strong robustness not only to common corruptions (as evaluated on ImageNet-C) but also to adversarial examples (ImageNet-A/ImageNet-P) and other distribution shifts like ImageNet-Sketch and Stylized-ImageNet, surpassing standard ViT models like DeiT and even robust CNNs in some cases[12].
- **Other ViT Variants:** The survey by Razzaq et al.[12] also mentions other ViT-related architectural approaches for robustness, such as Fully Attentional Networks (FAN) that consider both self-attention and channel attention, and replacing down-sampling blocks in CNNs with ViT-like patchifying input mechanisms.

Architectural innovations represent a more fundamental approach to robustness, aiming to embed resilience directly into the model's design. These methods often seek to improve how models process spatial information, handle shifts and aliasing, or adapt their internal parameters to changing input statistics, thereby reducing sensitivity to common corruptions. The interplay between architecture and training data (especially augmentations) is crucial, as robust architectures can better leverage diverse data to learn truly invariant features.

3.5.3 *Advanced Training Strategies.*

Beyond data augmentation and architectural modifications, advanced training strategies focus on altering the learning objective or the optimization process itself to guide models towards learning more robust and generalizable representations. These methods often aim to shape the loss landscape, enforce consistencies, or transfer knowledge from more robust sources.

Regularization Techniques: Regularization methods add constraints or penalties to the training process to prevent overfitting to the training data and encourage learning features that are less sensitive to input variations.

- **Adversarial Logit Pairing (ALP):** Originally developed by Kannan et al.[28] to improve robustness against adversarial examples, ALP adds a loss term that encourages the logits (pre-softmax outputs) of a clean image and its corresponding adversarial version to be similar. Interestingly, Hendrycks and Dietterich[25] found that an ALP defense, even when bypassed by stronger adversarial attacks, provided substantial robustness against common perturbations on their ImageNet-P benchmark. This suggests that enforcing similarity in the logit space for related inputs (even if one is adversarially perturbed) might lead to learning more stable underlying representations that also benefit common corruption robustness to some extent.
- **Sharpness-Aware Minimization (SAM):** Proposed by Foret et al.[17], SAM is a training procedure that seeks to find parameters that lie in ”flat” regions of the loss landscape, rather than sharp minima. It achieves this by minimizing the loss value not just at the current parameter point (w), but also the maximum loss within a small neighborhood around (w). The intuition is that models in flatter minima generalize better to unseen data and are less sensitive to small perturbations in either weights or inputs. While SAM was initially shown to improve generalization on clean data and robustness to label noise[17], subsequent research has explored its benefits for common corruption robustness. For example, Curvature Regularized SAM (CR-SAM)[59] explicitly incorporates a Hessian trace regularizer into SAM and demonstrates improved performance on ImageNet-C and ImageNet-R. FairSAM[9] adapts SAM to address fairness concerns under data corruption. The original SAM involves a two-step gradient computation, increasing training time[68], which has spurred research into more efficient variants like S2-SAM (Single-step SAM for Sparse training)[9] and RWP (Random Weight Perturbations)[32], though their direct impact on common corruption benchmarks like ImageNet-C is not always the primary focus of these efficiency-driven papers. However, the general principle of finding flatter minima is considered beneficial for overall robustness.
- **Other Regularization Methods:** Techniques like Jacobian Frequency Regularization (JaFR) by Chan et al.[4] aim to bias model learning towards specific frequency components by regularizing the Jacobian of the model. RoHL (Robustness via Hidden Layer Regularization) minimizes the total variation of activations in convolutional layers[12].

Contrastive and Self-Supervised Learning (SSL) for Robust Representations: Self-supervised learning methods, particularly those based on contrastive learning, have shown significant promise in learning representations that are inherently more robust to superficial input changes, including common corruptions[12].

- **Contrastive Learning Frameworks (e.g., SimCLR[5], MoCo):** These methods train models by maximizing the agreement (similarity) between different augmented ”views” of the same image (positive pairs) while minimizing agreement between views of different images (negative pairs) in an embedding space[12]. The augmentations used often include transformations that mimic common corruptions (e.g., color jitter, blur). By learning to be invariant to these augmentations, the learned representations tend to be more robust to similar corruptions encountered at test time.

- **Generalized Parametric Contrastive Learning (GPaCo):** Cui et al.[8] proposed GPaCo, a parametric contrastive learning approach that introduces learnable class-wise centers. It has demonstrated improved robustness on ImageNet-C, outperforming Masked Autoencoder (MAE) models, suggesting that the learned representations are more resilient to common corruptions.
- **Simple Data Mixing Prior (SDMP):** Ren et al.[43] proposed SDMP, which leverages data mixing techniques (like Mixup) within a self-supervised learning framework. It treats mixed samples as additional related views, enhancing representation learning and showing improved robustness on out-of-distribution samples.

Knowledge Distillation (KD) for Robustness Transfer: Knowledge distillation involves training a smaller "student" model to mimic the behavior of a larger, pre-trained, and often more robust "teacher" model[12]. This can be an effective way to transfer robustness.

- **NoisyStudent Training:** Developed by Xie et al.[62], this is an iterative self-training method. A teacher model, trained on labeled data, generates pseudo-labels for a large corpus of unlabeled images. A student model (of equal or greater capacity) is then trained on the combination of labeled and pseudo-labeled data, with significant noise (strong data augmentation like RandAugment, and model noise like dropout) applied during the student's training. The trained student then becomes the new teacher for the next iteration. NoisyStudent training has shown remarkable performance on ImageNet-C, demonstrating that leveraging large amounts of unlabeled data with appropriate noise and iterative refinement can lead to highly robust models.
- **Dataset Reinforcement:** Faghri et al.[15] proposed a strategy called Dataset Reinforcement, which combines knowledge distillation and data augmentation to create an "improved" version of a training dataset (e.g., ImageNet+). A strong teacher ensemble's knowledge is distilled onto various augmentations of the original dataset, and these "reinforced" examples are then used to train new models. This approach has shown significant robustness gains on ImageNet-R, ImageNet-A, and ImageNet-C.
- **Distilling Out-of-Distribution Robustness from Vision-Language Foundation Models:** Recent work[30] has focused on using large pre-trained vision-language models (VLMs) like CLIP as robust teachers. These models, trained on vast and diverse web-scale data, often exhibit impressive zero-shot generalization and robustness. The proposed Discrete Adversarial Distillation[72] (DAD) method uses a VQGAN to discretize adversarial examples generated from the VLM teacher, which are then used to train the student. This approach has yielded strong improvements on ImageNet-C and other OOD benchmarks.
- **Enhancing Weak Subnets (EWS):** Dubourg et al.[20] proposed EWS, a method that identifies "weak" sub-networks within a larger model that are particularly vulnerable to corruptions. These weak subnets are then explicitly strengthened via knowledge distillation from the full network, leading to improved overall corruption robustness. This technique is complementary to data augmentation methods.

These advanced training strategies often require careful tuning and can be computationally intensive, but they offer powerful ways to shape the learning process towards robustness. The success of methods like SAM, SSL, and KD highlights that how a model learns is as important as what data it sees or what its architecture is. The trend towards leveraging larger, more diverse datasets (even unlabeled) and distilling knowledge from powerful foundation models suggests a path towards more scalable and generalizable robustness. The common thread among many of these techniques is the encouragement of feature invariance to non-semantic transformations, a key characteristic for handling common image corruptions.

The broader landscape of robustness enhancement reveals a multifaceted approach. While data augmentation has been a dominant and evolving strategy, moving from simple transforms to complex learned and generative methods[12], its effectiveness is often amplified when combined with robust architectural designs[12] or sophisticated training paradigms[17]. This synergy underscores that no single solution is a panacea; rather, the most resilient models often emerge from a thoughtful integration of these complementary strategies. However, this progress is not without its costs. Many effective interventions introduce increased training complexity or computational overhead[7], presenting a "no free lunch" scenario where practitioners must balance the desired level of robustness against practical deployment constraints. The emergence of large foundation models and self-supervised learning as sources of inherent robustness signals a potential shift, where robustness might be more effectively "built-in" through large-scale pre-training on diverse data, rather than solely "added-on" to smaller models. Ultimately, the causal chain from diverse data exposure to learned feature invariance and finally to improved robustness appears to be a central theme underpinning many successful approaches.

4 DISCUSSION AND FUTURE DIRECTIONS

The extensive body of research surveyed in the preceding sections highlights significant progress in understanding and mitigating the impact of common image corruptions on deep learning models. However, it also reveals a complex landscape with persistent challenges and numerous avenues for future exploration.

4.1 Synthesis of Surveyed Works

The journey of research in common corruption robustness has seen a clear evolution. Initially, the focus was on quantifying the problem, largely spurred by the introduction of standardized benchmarks like ImageNet-C[25]. This led to an explosion of techniques, predominantly centered around data augmentation, which has proven to be a highly effective, albeit sometimes computationally intensive, strategy[12].

Another key observation is the inherent trade-offs associated with many robustness interventions. Enhanced robustness can sometimes come at the cost of a slight decrease in accuracy on clean, in-distribution data, although many modern techniques strive to mitigate this[26]. More significantly, there is often a trade-off with computational cost, both in terms of training time and model complexity[7]. Furthermore, robustness gained on a specific set of benchmark corruptions (e.g., those in ImageNet-C) does not always guarantee robustness to novel, unseen corruptions, highlighting the challenge of true out-of-distribution generalization[37].

The role of benchmarks themselves is a double-edged sword. While they have been crucial for standardizing evaluation and driving progress, there is a risk of ”teaching to the test” if research becomes overly focused on optimizing performance on a fixed set of corruptions[57]. The development of newer, more diverse benchmarks like ImageNet-R[24] and ImageNet-D[65] reflects an awareness of this limitation and a push towards more holistic evaluation.

Fundamentally, the problem of common corruption robustness is increasingly understood as a specific instance of the broader challenge of out-of-distribution (OOD) generalization[52]. Models fail because the statistical assumptions learned from the training distribution are violated by the corrupted inputs. This framing helps to connect research in corruption robustness with other areas of OOD generalization and transfer learning.

4.2 Identified Research Gaps, Unsolved Challenges, and Open Questions

Despite the progress, a primary research gap lies in achieving robust generalization to *unseen or novel corruptions.* Current methodologies are predominantly validated against predefined benchmark corruptions, such as those in ImageNet-C and its derivatives. However, real-world scenarios present a far more diverse and dynamic array of distortions, often involving combinations not explicitly captured by these benchmarks. Consequently, a significant challenge is the development of models that exhibit resilience to genuinely unanticipated corruptions, moving beyond robustness to a fixed list towards a more universal form of resilience [37].

This difficulty in generalizing to unseen corruptions is intrinsically linked to another critical gap: the lack of a deep *theoretical understanding of robustness.* While empirical successes are numerous, the fundamental principles governing why certain architectures, data augmentations, or training strategies effectively confer robustness often remain elusive [23]. Without a predictive theory to guide the design of new methods, the field largely relies on empirical trial and error. This reliance is particularly evident when considering *principled design versus empirical search* in areas like data augmentation, where many highly effective policies (e.g., AutoAugment [7]) are discovered through computationally intensive search procedures rather than being derived from a foundational understanding of robustness.

The empirical pursuit of robustness, while often effective, frequently introduces another challenge: achieving *efficient robustness.* State-of-the-art robustness can impose a considerable computational burden, whether during training, exemplified by methods like SAM’s double gradient computation [68] or complex augmentation pipelines, or in terms of model size. Developing methods that deliver strong robustness with minimal additional computational overhead and memory footprint is therefore crucial for practical deployment, especially on resource-constrained devices [53]. Furthermore, the challenge of robustness extends *beyond pixel-level corruptions*, which form the mainstay of current benchmarks like ImageNet-C. There is a pressing need to systematically investigate and enhance resilience to more structured or semantic corruptions, such as partial occlusions, significant geometric distortions (beyond simple affine transforms), adversarial patches, or alterations in object parts or context [41]. Current IQA-based metrics like VCR are also often limited to assessing pixel-level changes [48].

The limitations in current corruption types and evaluation methods naturally lead to questions about *reliable and comprehensive evaluation.* An ongoing debate queries whether existing benchmarks and metrics adequately capture the complexities of real-world robustness [57], with criticisms pointing to the lack of theoretical guidance in benchmark composition, arbitrary setting of perturbation levels, and potential disconnects between synthetic benchmark performance and real-world efficacy [57]. This necessitates the development of more benchmarks featuring genuine real-world corruptions, continuously varying severity levels, and evaluation metrics that better align with human perception and downstream task utility. Moreover, the focus of robustness research needs to expand to encompass *robustness in diverse tasks beyond classification.* While image classification has served as the primary testbed, other computer vision tasks, including object detection [37], semantic segmentation [41], person search [51], and video understanding [63], also demand robust models and may exhibit different sensitivities requiring tailored solutions. Underlying all these challenges is the need for a deeper *understanding and characterization of failure modes.* Analyzing precisely why and how models fail when confronted with specific corruptions, perhaps by examining shifts in internal representations or attention patterns [52], could yield invaluable insights for crafting more effective defenses.

5 CONCLUSION

This review has navigated the complex landscape of deep learning model robustness to common image distortions and corruptions, a critical attribute for reliable AI deployment. The journey from understanding the detrimental impact of these degradations to devising multifaceted mitigation strategies underscores a field of active and vital research.

The imperative for robust models is undeniable, driven by safety, reliability, and trust in AI systems operating in unpredictable environments. Common corruptions—spanning noise, blur, weather effects, and digital artifacts—can severely degrade the performance of even state-of-the-art deep learning models[46, 69]. Standardized benchmarks like `ImageNet-C`[25] have been instrumental in quantifying this problem and measuring progress.

A diverse array of techniques has been developed to enhance robustness. Data-centric approaches, particularly advanced augmentation methods like `AugMix`[26], have proven highly effective. Model-centric strategies involve designing inherently more robust architectures, such as modern CNNs like ConvNeXt[34] and specialized ViTs like FAN[73], alongside adaptable normalization layers[3] and frequency-aware designs[45]. Robust training methodologies, including specialized regularization[44] and dynamic sparse training[58], aim to instill resilience during the learning process. Furthermore, test-time adaptation (TTA) techniques[61, 70] offer practical means for models to adjust to corruptions encountered during inference.

Underpinning these advancements is a growing effort to understand the root causes of model fragility. Research into texture versus shape bias[22] and frequency-domain vulnerabilities[35] is shedding light on why models fail and guiding the development of more principled robustness solutions.

Despite significant strides, formidable challenges persist. Generalizing robustness to genuinely unseen and more complex real-world corruptions, managing the trade-offs between robustness, accuracy, and efficiency, and overcoming the limitations of current benchmarks are critical ongoing concerns. The interpretability of robustness mechanisms also remains a key area for deeper investigation.

Future research should prioritize the development of intrinsically robust architectures and training paradigms that foster generalization. Advanced data synthesis, more realistic and adaptive benchmarking, a stronger theoretical understanding of robustness, and strategies for lifelong continual adaptation are all crucial avenues. The ultimate aim is to build deep learning systems that not only achieve high accuracy on clean data but also maintain reliable performance and engender trust when faced with the diverse and unpredictable conditions of the real world.

REFERENCES

[1] Dan A Calian, Florian Stimberg, Olivia Wiles, Sylvestre-Alvise Rebuffi, Andras Gyorgy, Timothy Mann, and Sven Gowal. 2022. DEFENDING AGAINST IMAGE CORRUPTIONS THROUGH ADVERSARIAL AUGMENTATIONS. (2022).

[2] Elena Camuffo, Umberto Michieli, Simone Milani, Jijoong Moon, and Mete Ozay. 2024. Enhanced Model Robustness to Input Corruptions by Per-corruption Adaptation of Normalization Statistics. https://doi.org/10.48550/arXiv.2407.06450 arXiv:2407.06450 [cs].

[3] Elena Camuffo, Umberto Michieli, Jijoong Moon, Daehyun Kim, and Mete Ozay. 2024. FFT-based Selection and Optimization of Statistics for Robust Recognition of Severely Corrupted Images. https://doi.org/10.48550/arXiv.2403.14335 arXiv:2403.14335 [cs].

[4] Alvin Chan, Yew Soon Ong, and Clement Tan. 2022. How Does Frequency Bias Affect the Robustness of Neural Image Classifiers against Common Corruption and Adversarial Perturbations?. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Vienna, Austria, 659–665. https://doi.org/10.24963/ijcai.2022/93

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. arXiv:2002.05709 [cs.LG] https://arxiv.org/abs/2002.05709

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. https://doi.org/10.48550/arXiv.1604.01685 arXiv:1604.01685 [cs].

[7] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. 2019. AutoAugment: Learning Augmentation Policies from Data. https://doi.org/10.48550/arXiv.1805.09501 arXiv:1805.09501 [cs].

[8] Jiequan Cui, Zhisheng Zhong, Zhuotao Tian, Shu Liu, Bei Yu, and Jiaya Jia. 2023. Generalized Parametric Contrastive Learning. arXiv:2209.12400 [cs.CV] https://arxiv.org/abs/2209.12400

[9] Yucong Dai, Jie Ji, Xiaolong Ma, and Yongkai Wu. 2025. FairSAM: Fair Classification on Corrupted Data Through Sharpness-Aware Minimization. arXiv:2503.22934 [cs.LG] https://arxiv.org/abs/2503.22934

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848 ISSN: 1063-6919.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV] https://arxiv.org/abs/2010.11929

[12] Nathan Drenkow, Numair Sani, Ilya Shpitser, and Mathias Unberath. 2022. A Systematic Review of Robustness in Deep Learning for Computer Vision: Mind the gap? https://doi.org/10.48550/arXiv.2112.00639 arXiv:2112.00639 [cs].

[13] Shohei Enomoto, Naoya Hasegawa, Kazuki Adachi, Taku Sasaki, Shin'ya Yamaguchi, Satoshi Suzuki, and Takeharu Eda. 2024. Test-time Adaptation Meets Image Enhancement: Improving Accuracy via Uncertainty-aware Logit Switching. https://doi.org/10.48550/arXiv.2403.17423 arXiv:2403.17423 [cs].

[14] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* 88, 2 (June 2010), 303–338. https://doi.org/10.1007/s11263-009-0275-4

[15] Fartash Faghri, Hadi Pouransari, Sachin Mehta, Mehrdad Farajtabar, Ali Farhadi, Mohammad Rastegari, and Oncel Tuzel. 2023. Reinforce Data, Multiply Impact: Improved Model Accuracy and Robustness with Dataset Reinforcement. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Paris, France, 16986–16997. https://doi.org/10.1109/ICCV51070.2023.01562

[16] Jinyu Fan and Yi Zeng. 2023. Challenging deep learning models with image distortion based on the abutting grating illusion. *Patterns* 4, 3 (Feb. 2023), 100695. https://doi.org/10.1016/j.patter.2023.100695

[17] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2021. Sharpness-Aware Minimization for Efficiently Improving Generalization. https://doi.org/10.48550/arXiv.2010.01412 arXiv:2010.01412 [cs].

[18] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2022. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv:1811.12231 [cs.CV] https://arxiv.org/abs/1811.12231

[19] Gorana Gojić, Vladimir Vincan, Ognjen Kundačina, Dragiša Mišković, and Dinu Dragan. 2023. Non-adversarial Robustness of Deep Learning Methods for Computer Vision. https://doi.org/10.48550/arXiv.2305.14986 arXiv:2305.14986 [cs].

[20] Yong Guo, David Stutz, and Bernt Schiele. 2022. Improving Robustness by Enhancing Weak Subnets. https://doi.org/10.48550/arXiv.2201.12765 arXiv:2201.12765 [cs].

[21] Yong Guo, David Stutz, and Bernt Schiele. 2023. Improving Robustness of Vision Transformers by Reducing Sensitivity to Patch Corruptions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Vancouver, BC, Canada, 4108–4118. https://doi.org/10.1109/CVPR52729.2023.00400

[22] Edgar Heinert, Thomas Gottwald, Annika Mütze, and Matthias Rottmann. 2025. Shape Bias and Robustness Evaluation via Cue Decomposition for Image Classification and Segmentation. https://doi.org/10.48550/arXiv.2503.12453 arXiv:2503.12453 [cs].

[23] Markus Heinonen, Ba-Hien Tran, Michael Kampffmeyer, and Maurizio Filippone. 2025. Robust Classification by Coupling Data Mollification with Label Smoothing. https://doi.org/10.48550/arXiv.2406.01494 arXiv:2406.01494 [cs].

[24] 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. 2021. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Montreal, QC, Canada, 8320–8329. https://doi.org/10.1109/ICCV48922.2021.00823

[25] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. https://doi.org/10.48550/arXiv.1903.12261 arXiv:1903.12261 [cs].

[26] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2020. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. https://doi.org/10.48550/arXiv.1912.02781 arXiv:1912.02781 [stat].

[27] Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. https://doi.org/10.48550/arXiv.1902.09506 arXiv:1902.09506 [cs].

[28] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. 2018. Adversarial Logit Pairing. https://doi.org/10.48550/arXiv.1803.06373 arXiv:1803.06373 [cs].

[29] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 2022. 3D Common Corruptions and Data Augmentation. arXiv:2203.01441 [cs.CV] https://arxiv.org/abs/2203.01441

[30] Seetharam Killivalavan and Durairaj Thenmozhi. 2023. Software Metadata Classification based on Generative Artificial Intelligence. https://doi.org/10.48550/arXiv.2310.13006 arXiv:2310.13006 [cs].

[31] Alex Krizhevsky. [n. d.]. Learning Multiple Layers of Features from Tiny Images. ([n. d.]).

[32] Tao Li, Weihao Yan, Zehao Lei, Yingwen Wu, Kun Fang, Ming Yang, and Xiaolin Huang. 2022. Efficient Generalization Improvement Guided by Random Weight Perturbation. https://doi.org/10.48550/arXiv.2211.11489 arXiv:2211.11489 [cs].

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. https://doi.org/10.48550/arXiv.1405.0312 arXiv:1405.0312 [cs].

[34] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. arXiv:2201.03545 [cs.CV] https://arxiv.org/abs/2201.03545

[35] Harshitha Machiraju, Michael H. Herzog, and Pascal Frossard. 2023. Frequency-Based Vulnerability Analysis of Deep Learning Models against Image Corruptions. https://doi.org/10.48550/arXiv.2306.07178 arXiv:2306.07178 [cs].

[36] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. 2022. Towards Robust Vision Transformer. https://doi.org/10.48550/arXiv.2105.07926 arXiv:2105.07926 [cs].

[37] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. [n. d.]. Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming. ([n. d.]).

[38] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. 2020. Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming. https://doi.org/10.48550/arXiv.1907.07484 arXiv:1907.07484 [cs].

[39] Eric Mintun, Alexander Kirillov, and Saining Xie. 2021. On Interaction Between Augmentations and Corruptions in Natural Corruption Robustness. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., 3571–3583. https://proceedings.neurips.cc/paper_files/paper/2021/hash/1d49780520898fe37f0cd6b41c5311bf-Abstract.html

[40] Apostolos Modas, Rahul Rade, Guillermo Ortiz-Jiménez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. 2022. PRIME: A few primitives can boost robustness to common corruptions. https://doi.org/10.48550/arXiv.2112.13547 arXiv:2112.13547 [cs].

[41] Seif Mzoughi, Mohamed Elshafeia, and Foutse Khomh. 2025. Evaluating and Enhancing Segmentation Model Robustness with Metamorphic Testing. https://doi.org/10.48550/arXiv.2504.02335 arXiv:2504.02335 [cs].

[42] Keiron O'Shea and Ryan Nash. 2015. An Introduction to Convolutional Neural Networks. arXiv:1511.08458 [cs.NE] https://arxiv.org/abs/1511.08458

[43] Sucheng Ren, Huiyu Wang, Zhengqi Gao, Shengfeng He, Alan Yuille, Yuyin Zhou, and Cihang Xie. 2022. A Simple Data Mixing Prior for Improving Self-Supervised Learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New Orleans, LA, USA, 14575–14584. https://doi.org/10.1109/CVPR52688.2022.01419

[44] Adrián Rodríguez-Muñoz, Tongzhou Wang, and Antonio Torralba. 2024. Characterizing Model Robustness via Natural Input Gradients. https://doi.org/10.48550/arXiv.2409.20139 arXiv:2409.20139 [cs].

[45] Tonmoy Saikia, Cordelia Schmid, and Thomas Brox. 2021. Improving robustness against common corruptions with frequency biased models. https://doi.org/10.48550/arXiv.2103.16241 arXiv:2103.16241 [cs].

[46] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. 2020. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 11539–11551. https://proceedings.neurips.cc/paper/2020/hash/85690f81aadc1749175c187784afc9ee-Abstract.html

[47] George Shen. [n. d.]. Analyzing the Effects of Data Corruptions on Machine Learning Mod- els. ([n. d.]).

[48] Huakun Shen, Boyue Hu, Krzysztof Czarnecki, Lina Marsso, and Marsha Chechik. [n. d.]. Assessing Visually-Continuous Corruption Robustness of Neural Networks Relative to Human Performance. ([n. d.]).

[49] Yu Shen, Laura Zheng, Manli Shu, Weizi Li, Tom Goldstein, and Ming Lin. 2021. Gradient-Free Adversarial Training Against Image Corruption for Learning-based Steering. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., 26250–26263. https://proceedings.neurips.cc/paper/2021/hash/dce8af15f064d1accb98887a21029b08-Abstract.html

[50] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA Models That Can Read. https://doi.org/10.48550/arXiv.1904.08920 arXiv:1904.08920 [cs].

[51] Woojung Son, Yoonki Cho, Guoyuan An, Jinhwan Seo, Chanmi Lee, and Sung-eui Yoon. 2024. Towards Robustness of Person Search against Corruptions. (Oct. 2024). https://openreview.net/forum?id=gY2IHLUJhk

[52] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. 2020. Measuring Robustness to Natural Distribution Shifts in Image Classification. https://doi.org/10.48550/arXiv.2007.00644 arXiv:2007.00644 [cs].

[53] Trung Trinh, Markus Heinonen, Luigi Acerbi, and Samuel Kaski. 2024. Improving robustness to corruptions with multiplicative weight perturbations. https://doi.org/10.48550/arXiv.2406.16540 arXiv:2406.16540 [cs].

[54] Muhammad Usama, Syeda Aishah Asim, Syed Bilal Ali, Syed Talal Wasim, and Umair Bin Mansoor. 2025. Analysing the Robustness of Vision-Language-Models to Common Corruptions. arXiv:2504.13690 [cs.CV] https://arxiv.org/abs/2504.13690

[55] Muhammad Usama, Syeda Aishah Asim, Syed Bilal Ali, Syed Talal Wasim, and Umair Bin Mansoor. 2025. Analysing the Robustness of Vision-Language-Models to Common Corruptions. https://doi.org/10.48550/arXiv.2504.13690 arXiv:2504.13690 [cs].

[56] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. 2018. Rotation Equivariant CNNs for Digital Pathology. (June 2018). arXiv:1806.03962 [cs.CV]

[57] Shunxin Wang, Raymond Veldhuis, Christoph Brune, and Nicola Strisciuglio. 2024. A Survey on the Robustness of Computer Vision Models against Common Corruptions. https://doi.org/10.48550/arXiv.2305.06024 arXiv:2305.06024 [cs].

[58] Boqian Wu, Qiao Xiao, Shunxin Wang, Nicola Strisciuglio, Mykola Pechenizkiy, Maurice van Keulen, Decebal Constantin Mocanu, and Elena Mocanu. 2025. DYNAMIC SPARSE TRAINING VERSUS DENSE TRAIN- ING: THE UNEXPECTED WINNER IN IMAGE CORRUP- TION ROBUSTNESS. (2025).

[59] Tao Wu, Tie Luo, and Donald C. Wunsch. 2023. CR-SAM: Curvature Regularized Sharpness-Aware Minimization. https://doi.org/10.48550/arXiv.2312.13555 arXiv:2312.13555 [cs].

[60] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. [n. d.]. End-to-End Deep Learning for Person Search. ([n. d.]).

[61] Zehao Xiao and Cees G. M. Snoek. 2024. Beyond Model Adaptation at Test Time: A Survey. https://doi.org/10.48550/arXiv.2411.03687 arXiv:2411.03687 [cs].

[62] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. 2020. Self-training with Noisy Student improves ImageNet classification. https://doi.org/10.48550/arXiv.1911.04252 arXiv:1911.04252 [cs].

[63] Chenyu Yi, Siyuan Yang, Yufei Wang, Haoliang Li, Yap-Peng Tan, and Alex C. Kot. 2023. Temporal Coherent Test-Time Optimization for Robust Video Classification. https://doi.org/10.48550/arXiv.2302.14309 arXiv:2302.14309 [cs].

[64] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. https://doi.org/10.48550/arXiv.1905.04899 arXiv:1905.04899 [cs].

[65] Chenshuang Zhang, Fei Pan, Junmo Kim, In So Kweon, and Chengzhi Mao. 2024. ImageNet-D: Benchmarking Neural Network Robustness on Diffusion Synthetic Object. https://doi.org/10.48550/arXiv.2403.18775 arXiv:2403.18775 [cs].

[66] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. https://doi.org/10.48550/arXiv.1710.09412 arXiv:1710.09412 [cs].

[67] Richard Zhang. 2019. Making Convolutional Networks Shift-Invariant Again. https://doi.org/10.48550/arXiv.1904.11486 arXiv:1904.11486 [cs].

[68] Yihao Zhang, Hangzhou He, Jingyu Zhu, Huanran Chen, Yifei Wang, and Zeming Wei. 2024. On the Duality Between Sharpness-Aware Minimization and Adversarial Training. https://doi.org/10.48550/arXiv.2402.15152 arXiv:2402.15152 [cs].

[69] Yunlong Zhang, Yuxuan Sun, Honglin Li, Sunyi Zheng, Chenglu Zhu, and Lin Yang. 2022. Benchmarking the Robustness of Deep Neural Networks to Common Corruptions in Digital Pathology. https://doi.org/10.48550/arXiv.2206.14973 arXiv:2206.14973 [cs].

[70] Yufei Zhang, Yicheng Xu, Hongxin Wei, Zhiping Lin, and Huiping Zhuang. 2024. Analytic Continual Test-Time Adaptation for Multi-Modality Corruption. https://doi.org/10.48550/arXiv.2410.22373 arXiv:2410.22373 [cs].

[71] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. 2017. Person Re-identification in the Wild. https://doi.org/10.48550/arXiv.1604.02531 arXiv:1604.02531 [cs].

[72] Andy Zhou, Jindong Wang, Yu-Xiong Wang, and Haohan Wang. 2023. Distilling Out-of-Distribution Robustness from Vision-Language Foundation Models. *Advances in Neural Information Processing Systems* 36 (Dec. 2023), 32938–32957. https://proceedings.neurips.cc/paper_files/paper/2023/hash/67f30132d98e758f7b4e28c36091d86e-Abstract-Conference.html

[73] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M. Alvarez. 2022. Understanding The Robustness in Vision Transformers. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 27378–27394. https://proceedings.mlr.press/v162/zhou22m.html ISSN: 2640-3498.