When Do Deep Ensembles Improve Robustness

to Spurious Correlations?

CSE3000 Research Project

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 25, 2026

Name of the student: Jaouad Hidayat
Final project course: CSE3000 Research Project
Thesis committee: Wendelin Böhmer (Responsible Professor, Supervisor), David Tax (Examiner)

WHEN DO DEEP ENSEMBLES IMPROVE ROBUSTNESS TO SPURIOUS CORRELATIONS?

Jaouad Hidayat
Delft University of Technology

ABSTRACT

Models trained with empirical risk minimization can rely on spurious features that are highly predictive during training but fail under distribution shift. We study deep ensembles as a simple baseline that does not require spurious-attribute labels. We construct a controlled dataset by placing MNIST digits on CIFAR-10 backgrounds. During training, each digit class is assigned its own disjoint set of N backgrounds, so background identity alone predicts the label. We evaluate in-distribution (ID) and two out-of-distribution (OOD) shifts: (i) *seen-shuffle*, which keeps the same backgrounds but permutes their association with labels, and (ii) *unseen-background*, which draws backgrounds from a held-out pool and therefore combines shortcut breaking with a novel-background shift. Across $N \in \{1, 2, 4, 8, 16, 32, 64\}$, deep ensembles improve OOD accuracy most in an intermediate regime (max gain $17.1\,\mathrm{pp}$ at $N{=}8$ for $M{=}8$; 95% CI $8.4\,\mathrm{pp}$ to $26.8\,\mathrm{pp}$). For LeNet trained with mean-squared error (MSE), increasing ensemble size from $M{=}1$ to $M{=}8$ improves *seen-shuffle* OOD accuracy from $71.3\,\%$ to $88.4\,\%$ at $N{=}8$. In the same setting, the ensemble prediction is less aligned with the background label in seen-shuffle: background-follow rate drops from $26.9\,\%$ ($M{=}1$) to $16.3\,\%$ ($M{=}8$) at $N{=}8$. Under the unseen-background shift, we observe larger gains at smaller N (for example, $45.7\,\%$ to $67.6\,\%$ at $N{=}4$), but this setting changes both the shortcut and the background distribution. We also find that model capacity and loss choice affect where shortcut reliance breaks down, and that a parameter-matched larger single model can outperform a small ensemble. Overall, deep ensembles can improve OOD accuracy and reduce shortcut alignment in this controlled setup, but the effect is regime-dependent and should be interpreted relative to the specific shift.

1 INTRODUCTION

Standard empirical risk minimization often exploits shortcuts. When multiple features predict the label, a network can lock onto a feature that is easy to fit but unstable under distribution shift (Geirhos et al., 2020). In vision, background is a recurring shortcut. A classifier can learn "cow" from grass rather than from the animal, and fail when the background changes. This is a modern form of the Clever Hans effect (Ye et al., 2024; Lapuschkin et al., 2019).

Many approaches target spurious correlations by adding structure. Group distributionally robust optimization improves worst-group performance but relies on group labels (Sagawa et al., 2020). When group labels are missing, methods such as environment inference or two-stage reweighting can help but introduce additional choices and tuning (Creager et al., 2021; Liu et al., 2021). Since shortcuts are common in practice (Koh et al., 2021), it is useful to understand how far simple baselines can go.

This paper studies deep ensembles. Deep ensembles are a strong baseline for predictive uncertainty (Lakshminarayanan et al., 2017) and can be competitive under dataset shift (Ovadia et al., 2019). Our goal is narrower and empirical: we isolate a deterministic background shortcut and ask when ensembling improves *OOD accuracy under a specified shift* and whether those improvements coincide with reduced shortcut following.

Research question: When does increasing ensemble size M improve OOD robustness in the presence of a deterministic spurious background–label shortcut, and does it reduce direct measures of shortcut reliance?

Contributions: We (1) evaluate deep ensembles on a controllable MNIST-on-CIFAR shortcut benchmark under two OOD shifts (seen-shuffle and unseen-background), (2) add a direct shortcut metric (background-follow rate) for seen-shuffle to avoid equating OOD accuracy with shortcut avoidance, and (3) quantify how the regime where ensembling helps depends on loss and model capacity, including a comparison to a parameter-matched larger single model (Abe et al., 2022).

Paper structure: Section 2 discusses related work, Section 3 defines the dataset and evaluation shifts, Section 4 details the experimental protocol, and Sections 5–6 present results and limitations.

Figure 1 gives a visual overview of the synthetic data construction and the evaluation splits (illustration uses $N = 2$).

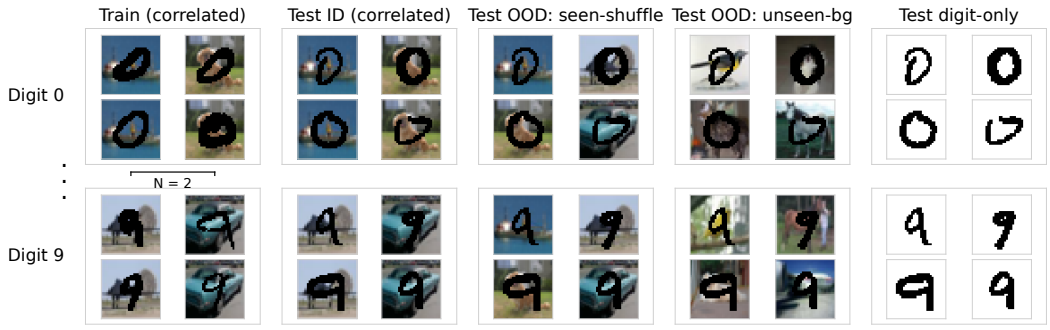

Figure 1: Dataset and evaluation modes (illustration uses $N = 2$). Training and ID test draw backgrounds from class-specific sets $\{\mathcal{B}_y\}$, which are disjoint across digit classes. Seen-shuffle OOD keeps the same backgrounds but permutes their association with labels, isolating shortcut reliance without introducing novel backgrounds. Unseen-background OOD draws from a held-out background pool, breaking the shortcut but also introducing a background-identity shift. The digit-only split removes backgrounds and serves as an approximate ceiling for digit-based recognition.

2 Related Work

Shortcut learning and spurious correlations are widely documented (Geirhos et al., 2020; Ye et al., 2024). Canonical spurious-correlation benchmarks include Colored MNIST (Arjovsky et al., 2019) and object-background confounding datasets such as Waterbirds (Sagawa et al., 2020). In such settings, methods that use group information can substantially improve worst-group accuracy (Sagawa et al., 2020). When group labels are unavailable, approaches such as environment inference (Creager et al., 2021) and Just Train Twice (Liu et al., 2021) aim to recover robustness with minimal supervision.

Beyond these methods, a large literature studies label-free or weakly supervised debiasing and domain generalization; for example, Learning from Failure (Nam et al., 2020) and risk extrapolation (Krueger et al., 2021). Background dependence in vision models has also been analyzed directly in more realistic settings (Xiao et al., 2020).

Deep ensembles are a simple and effective baseline for uncertainty estimation (Lakshminarayanan et al., 2017). Large-scale evaluations under dataset shift find ensembles to be competitive for both accuracy and uncertainty (Ovadia et al., 2019). A complementary question is whether an ensemble provides benefits beyond simply using a larger single model (Abe et al., 2022). Our controlled benchmark varies N, the number of backgrounds associated with each label, which controls the complexity of the background–label shortcut and lets us probe when ensembling helps under different evaluation shifts.

We focus on independently trained ensembles; other "cheap" ensemble-like baselines (e.g., snapshot ensembles (Huang et al., 2017) or stochastic weight averaging (Izmailov et al., 2018)) are relevant alternatives but are outside our experimental scope.

3 METHODOLOGY

3.1 DATASET CONSTRUCTION

We build a synthetic shortcut dataset from MNIST (LeCun et al., 1998) digits and CIFAR-10 (Krizhevsky, 2009) backgrounds. Each MNIST digit is padded from 28×28 to 32×32 and converted to a binary mask. Pixels under the digit are set to zero in the CIFAR background, yielding a black digit silhouette on a natural background.

Class-specific background world: For each dataset instance and each digit class $y \in \{0, \ldots, 9\}$, we sample a disjoint set of N CIFAR-10 images as backgrounds, denoted $\mathcal{B}_y$ (CIFAR labels are not used). During training, each digit example with label y is paired with a background sampled from $\mathcal{B}_y$. The background index for each example is fixed when the dataset is generated (sampling from $\mathcal{B}_y$ with replacement), so backgrounds are not resampled across epochs. Since the sets are disjoint, background identity alone predicts y in the training distribution, so a background–label shortcut exists by construction. Increasing N increases the number of backgrounds associated with each label and makes this shortcut harder to fit for a fixed model and training budget.

Input-mode control: To verify that the digit signal itself is sufficient, we also train a *digit-only* control where the background is removed (input is the masked digit only).

3.2 EVALUATION SHIFTS (ID AND TWO OOD MODES)

We evaluate ID on a correlated test set that preserves the background-to-label mapping.

We use two OOD modes:

1. **Seen-shuffle OOD:** test examples use backgrounds sampled from the same sets $\{\mathcal{B}_y\}$, but the association between background-set and label is permuted. Backgrounds are therefore in-distribution, but systematically misleading if the model relies on them.
2. **Unseen-background OOD:** test examples use backgrounds drawn from a held-out CIFAR-10 pool that is disjoint from all backgrounds used in the class-specific world. This held-out pool is reserved at dataset generation time and is never used in training or ID evaluation. This breaks the shortcut but also introduces a novel-background shift.

Background-follow rate: In the seen-shuffle setting, each example has a well-defined "background class" y_{bg} (the class whose background set supplied the image). We measure shortcut alignment of the *final predictor* with the background label using the background-follow rate:

$$\mathrm{BFR} = \mathbb{E}\big[\mathbf{1}\{\hat{y}(x) = y_{\mathrm{bg}}\}\big], \tag{1}$$

where $\hat{y}(x) = \arg\max_y \bar{p}(y \mid x)$ is the predicted label (for single models, $\bar{p} = p_1$). A high BFR indicates that predictions track the background label in this shift. Because BFR is computed after probability averaging, a reduction in BFR should be interpreted as reduced *background alignment of the ensemble predictor*, not necessarily as a direct measurement of per-member feature usage.

3.3 MODELS, TRAINING, AND ENSEMBLES

We study a LeNet-style CNN (LeCun et al., 1998) (about 62k parameters) and Wide ResNets (WRN-16-1 and WRN-28-1) (Zagoruyko & Komodakis, 2016) (about 175k and 369k parameters, respectively). We train for 30 epochs with SGD (momentum 0.9), batch size 256, learning rate 1.0 (constant), and no weight decay or dropout. Unless stated otherwise, digits are rendered with the mask procedure above and spurious strength is deterministic (background fully predicts label in training).

Table 1: Key quantities used in the results. N_{50} and N_{90} denote the smallest N where single-model *seen-shuffle* OOD accuracy exceeds $50\,\%$ and $90\,\%$, respectively. "Max gain" is the maximum OOD improvement from ensembling (relative to $M{=}1$) within the evaluated M range. Confidence intervals are 95% hierarchical bootstrap intervals.

Setting	Params	N_{50}	N_{90}	Key effect
LeNet seen-shuffle (MSE)	62k	8	32	Max gain: $17.1\,\mathrm{pp}$ at $N = 8$ (95% CI $8.4\,\mathrm{pp}$ to $26.8\,\mathrm{pp}$); BFR at $N = 8$ drops from $26.9\,\%$ ($M = 1$) to $16.3\,\%$ ($M = 8$).
LeNet unseen-background (MSE)	62k	8	32	Max gain: $21.9\,\mathrm{pp}$ at $N = 4$ (95% CI $12.0\,\mathrm{pp}$ to $30.8\,\mathrm{pp}$).
LeNet seen-shuffle (CE)	62k	16	32	Max gain: $5.3\,\mathrm{pp}$ at $N = 16$ (95% CI $0.3\,\mathrm{pp}$ to $9.9\,\mathrm{pp}$).
WRN16 seen-shuffle (MSE, filtered)	175k	16	64	Max gain: $2.5\,\mathrm{pp}$ at $N = 16$ (95% CI $-0.8\,\mathrm{pp}$ to $5.9\,\mathrm{pp}$).
WRN28 seen-shuffle (MSE, filtered)	369k	16	32	Single-model only ($M = 1$); used for parameter-matched comparison.

We consider ensembles of size M by training M independent members with different random seeds. To keep the protocol consistent across M, each run trains the maximum ensemble size and smaller M are evaluated as prefixes. The ensemble predictive distribution averages member probabilities:

$$\bar{p}(y \mid x) = \frac{1}{M} \sum_{m=1}^{M} p_m(y \mid x). \tag{2}$$

3.4 Aggregation and confidence intervals

For each configuration we average across independent dataset seeds and training seeds. We report 95% confidence intervals computed with a hierarchical bootstrap that resamples dataset seeds (outer level) and training seeds within each dataset (inner level).

4 Experiments

We sweep backgrounds-per-class $N \in \{1, 2, 4, 8, 16, 32, 64\}$. For LeNet, the main evaluation uses **seen-shuffle OOD** with $M \in \{1, 2, 4, 8\}$, and we additionally report **unseen-background OOD** with the same M values as a second, qualitatively different shift (plus a digit-only $M{=}1$ control). We also run a **loss ablation** (MSE versus cross-entropy) for LeNet under seen-shuffle OOD. For capacity comparisons, we evaluate WRN-16-1 with $M \in \{1, 2\}$ and WRN-28-1 with $M{=}1$ under seen-shuffle OOD.

For WRN models, a subset of runs does not fit the correlated test distribution (unusually low correlated-test accuracy). To keep architecture comparisons interpretable, we apply a conservative filter and exclude WRN rows with correlated-test accuracy below 0.90. Appendix Table 3 reports excluded counts per N.

Unless stated otherwise, each LeNet point averages 5 independently generated datasets and 4 training initializations per dataset (20 runs per N, M), and each WRN point averages 5 datasets and 2 training initializations per dataset (10 runs per N, M).

Table 1 lists the key summary quantities we use throughout the results.

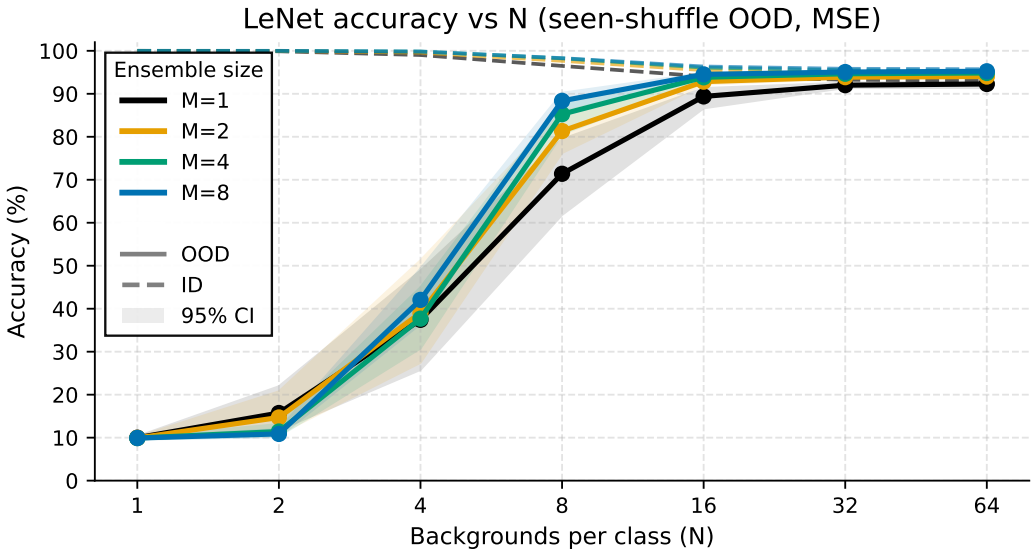

Figure 2: LeNet, MSE loss, seen-shuffle OOD. Solid lines show OOD accuracy; dashed lines show ID accuracy on the correlated test set. Shaded regions are 95% bootstrap confidence intervals. Ensemble gains peak at intermediate N ($N \approx 8$).

5 Results

5.1 Main result: ensembles improve seen-shuffle OOD accuracy at intermediate N

Figure 2 shows LeNet accuracy under **seen-shuffle** OOD. ID accuracy remains high across settings, while OOD accuracy depends strongly on N and M. Ensembling helps most when the shortcut is neither trivial nor already too weak. For example, at $N{=}8$ increasing M from 1 to 8 raises OOD accuracy from $71.3\,\%$ to $88.4\,\%$ (a $17.1\,\mathrm{pp}$ gain; 95% CI $8.4\,\mathrm{pp}$ to $26.8\,\mathrm{pp}$). At $N{=}4$, gains are small ($37.5\,\%$ to $42.1\,\%$), suggesting that this shift can be too adversarial for probability averaging to help much. At larger N (e.g., $N \geq 32$), single-model OOD accuracy is already above $91\,\%$ and gains from ensembling are small. Unless stated otherwise, results in this section use MSE; Section 5.4 shows that switching to cross-entropy changes both the transition point in N and the magnitude of ensemble gains.

5.2 Seen-shuffle isolates shortcut following, and ensemble predictions reduce shortcut alignment at $N{=}8$

Seen-shuffle OOD is designed to make the background cue *misleading while remaining in-distribution*. To separate shortcut following from effects of changing the background distribution, Figure 3 compares seen-shuffle OOD to unseen-background OOD and reports background-follow rate (BFR).

Two patterns stand out. First, for small N, BFR is near $100\,\%$, indicating heavy shortcut reliance; ensembling does not mitigate this regime. Second, at $N{=}8$, the ensemble reduces BFR from $26.9\,\%$ ($M{=}1$) to $16.3\,\%$ ($M{=}8$), while improving seen-shuffle OOD accuracy from $71.3\,\%$ to $88.4\,\%$. This is consistent with the OOD improvements in this regime coinciding with reduced shortcut *alignment* in the ensemble prediction (as measured by BFR). However, because BFR is computed on the final averaged predictor, it does not by itself distinguish whether individual members rely less on the background or whether averaging cancels background-driven disagreements.

At $N{=}4$ and $M{=}8$, seen-shuffle yields lower accuracy than unseen-background ($42.1\,\%$ versus $67.6\,\%$; difference $25.5\,\mathrm{pp}$, 95% CI $22.6\,\mathrm{pp}$ to $29.0\,\mathrm{pp}$). One interpretation is that a misleading in-distribution shortcut can be harder to overcome than the absence of the shortcut cue, but these shifts are not directly comparable because unseen-background also changes the background distribution.

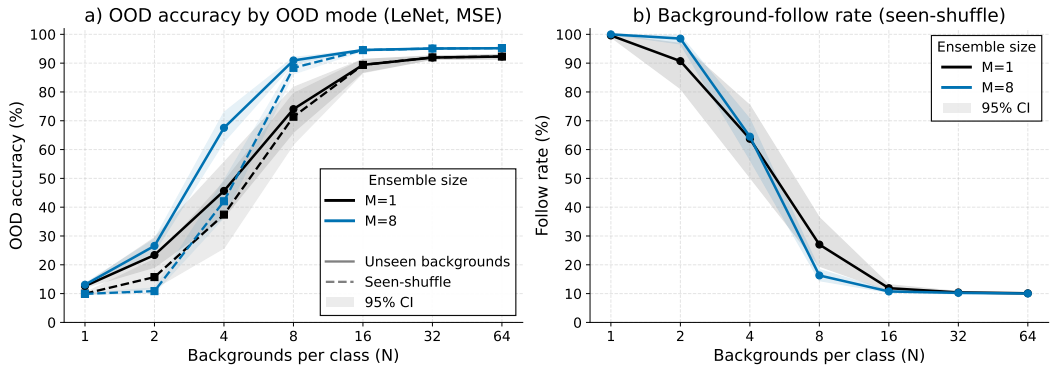

Figure 3: LeNet, MSE. (a) OOD accuracy under two OOD shifts. (b) Background-follow rate (BFR) for seen-shuffle OOD. High BFR indicates strong alignment with the background label in this shift.

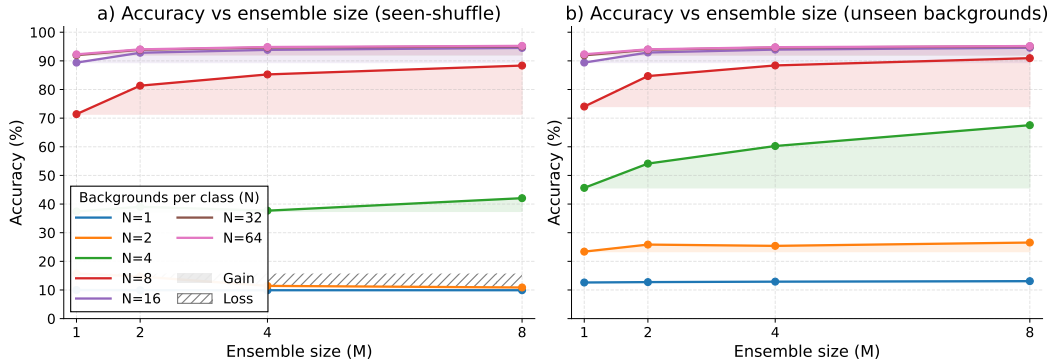

Figure 4: OOD accuracy versus ensemble size M (LeNet, MSE). Left: seen-shuffle OOD (main). Right: unseen-background OOD (secondary shift). Improvements are N-dependent and show diminishing returns in M.

5.3 Ensemble size interacts with N and OOD mode

Figure 4 plots OOD accuracy as a function of ensemble size M for selected N. When the shortcut is overwhelming (small N), increasing M has little effect. When the task is already robust (large N), gains are small. The largest improvements occur at intermediate N, and appear early: for seen-shuffle at $N{=}8$, most of the gain occurs by $M{=}2$ and then saturates. The unseen-background shift exhibits a different peak, with larger gains already at $N{=}4$.

5.4 Loss choice matters in the shortcut regime

Figure 5 compares mean-squared error (MSE) and cross-entropy (CE) for LeNet under **seen-shuffle** OOD. CE performs similarly at large N, but MSE is substantially better in the intermediate regime where shortcut reliance is strongest. At $N{=}8$, the single-model gap is $33.1\,\mathrm{pp}$ in favor of MSE (95% CI $25.0\,\mathrm{pp}$ to $40.3\,\mathrm{pp}$), and for $M{=}8$ the gap grows to $51.1\,\mathrm{pp}$ (95% CI $48.7\,\mathrm{pp}$ to $53.9\,\mathrm{pp}$). At $N{=}64$, CE is slightly better (MSE $-$ CE $-2.7\,\mathrm{pp}$; 95% CI $-2.9\,\mathrm{pp}$ to $-2.5\,\mathrm{pp}$). We therefore use MSE as the default for the main ensemble analysis and report CE as a controlled ablation.

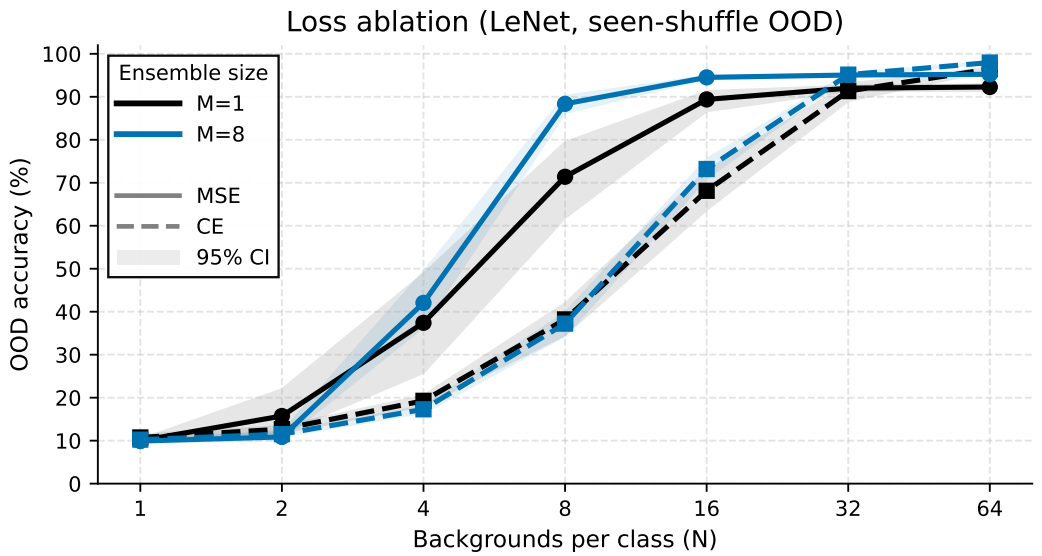

Figure 5: Loss ablation for LeNet under seen-shuffle OOD. MSE dramatically improves robustness in the intermediate-N regime (especially near $N = 8$), while CE matches or slightly exceeds MSE once single models are already robust.

5.5 Model capacity shifts the transition away from shortcut reliance

Figure 6 compares single-model ($M{=}1$) **seen-shuffle** OOD accuracy across architectures and includes the digit-only control. Digit-only LeNet achieves about $94\,\%$ OOD accuracy even at $N{=}1$, showing that digit information is sufficient for good OOD performance, and that failures in the composite setting are not due to the masking procedure alone.

In contrast, larger models remain non-robust at higher N. At $N{=}8$, single-model LeNet reaches $71.3\,\%$ OOD accuracy, while WRN-16-1 is at $13.0\,\%$ and WRN-28-1 at $26.6\,\%$. The breakpoint view in Table 1 shows the same pattern: WRN-16-1 requires $N{=}64$ to exceed $90\,\%$ OOD accuracy, whereas LeNet reaches that level at $N{=}32$. This is consistent with a "lookup table" picture: higher-capacity models can fit the background-to-label mapping for larger N, delaying the transition toward digit-based features. However, this comparison is conditional on our training recipe and on the WRN stability filter (Appendix A); we therefore treat it as evidence about this benchmark rather than as a general claim about capacity. For WRN-16-1 specifically, the observed $M{=}2$ ensemble gain over a single model is small and not statistically distinguishable from zero within our sample size (Table 1).

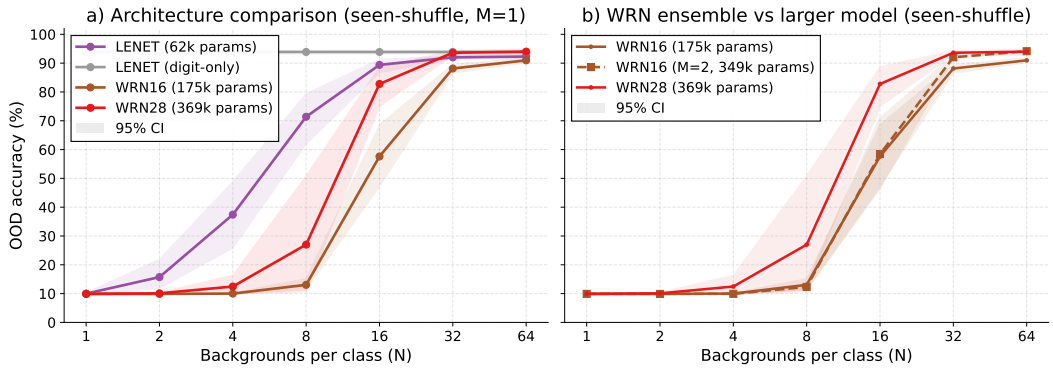

Figure 6: Capacity and scaling under seen-shuffle OOD. (a) Single-model comparison across architectures, including a digit-only control (approximate ceiling). (b) Parameter-matched comparison: WRN-16-1 with $M{=}2$ (349k params) versus WRN-28-1 (369k params). The larger single model is substantially better at $N{=}16$ (difference $25.4\,\text{pp}$, 95% CI $7.4\,\text{pp}$ to $42.5\,\text{pp}$), while the gap narrows once both reach high accuracy.

6 Discussion and Limitations

Our results support three main points. First, deep ensembles can substantially improve OOD accuracy in a regime where the background shortcut is learnable but not perfectly stable across training runs. Second, when we isolate shortcut following with the seen-shuffle OOD mode, ensemble improvements at $N{=}8$ coincide with a clear reduction in background-follow rate for the *ensemble prediction*. Third, the transition away from shortcut reliance depends on model capacity and loss, and a parameter-matched larger model can outperform a small ensemble.

Several limitations remain. The benchmark is synthetic, and we do not validate on standard real-world spurious-correlation datasets such as Waterbirds or WILDS benchmarks (Koh et al., 2021). Our shortcut is deterministic (spurious strength 1.0), while real spurious correlations are often partial. Unseen-background OOD mixes shortcut breaking with a novel-background shift, so conclusions about shortcut alignment should rely primarily on the seen-shuffle analysis and the BFR metric. BFR captures only the failure mode of predicting the background class; a model may still use background features without predicting y_{bg}. Our main results use MSE, and the loss ablation (Section 5.4) shows that conclusions about when ensembling helps are sensitive to this choice. For WRN models, we apply a stability filter (Appendix A), which improves interpretability but may bias architecture comparisons. Finally, we only evaluate ensembles up to $M{=}8$ for LeNet and $M{=}2$ for WRN-16-1, so we cannot characterize scaling beyond these sizes or compare to alternative ensemble-like baselines such as snapshot ensembles or SWA (Huang et al., 2017; Izmailov et al., 2018).

Threats to validity:

- **Internal validity:** conclusions about "when ensembling helps" are conditional on our fixed training recipe, and for WRNs additionally on the correlated-test stability filter (Appendix A).
- **Construct validity:** BFR operationalizes one specific notion of shortcut following (predicting the background class) and should not be read as a complete attribution of background reliance.
- **External validity:** results are obtained on a synthetic benchmark with deterministic spurious strength, and may not transfer directly to real spurious-correlation datasets without additional validation.

7 CONCLUSION

We investigated when deep ensembles improve OOD accuracy under a controlled spurious background–label shortcut. In a MNIST-on-CIFAR benchmark with disjoint class-specific background sets, ensemble gains are largest at intermediate N, where single models are neither hopelessly shortcut-driven nor already robust. Using seen-shuffle OOD and a direct shortcut metric, we find that at $N{=}8$ a LeNet $M{=}8$ ensemble (trained with MSE) both improves OOD accuracy ($71.3\,\%$ to $88.4\,\%$) and reduces background-follow rate ($26.9\,\%$ to $16.3\,\%$) relative to a single model. We also show that larger architectures can rely on the shortcut at higher N, shifting where ensembling is beneficial, and that parameter-matched larger single models can outperform small ensembles. Overall, deep ensembles are a strong label-free baseline in this controlled setting, but their benefits depend on the shortcut regime, the chosen shift, and design choices such as the loss.

8 RESPONSIBLE RESEARCH

8.1 REPRODUCIBILITY AND INTEGRITY

All experiments use fixed, logged random seeds for dataset construction and training. We separate randomness from dataset generation (outer seed) and optimization (inner seed). Reported confidence intervals are computed with a hierarchical bootstrap that respects this nesting. Smaller-M results are evaluated as prefixes of the maximum ensemble size within each run; Appendix A describes the protocol and reports a subset-variance diagnostic.

8.2 COMPUTE AND ENVIRONMENTAL COST

Across all runs we trained 3,710 network instances (counting ensemble members) for 30 epochs each, with maximum ensemble sizes $M{=}8$ (LeNet) and $M{=}2$ (WRN-16-1). Training ran on a single NVIDIA GeForce RTX 4070 Laptop GPU and took 17.1 GPU-hours (wall-clock). Following standard footprint reporting practice (Henderson et al., 2020), we estimate operational energy as $E = P\,t$. Assuming the GPU operated at its maximum subsystem power ($115\,\mathrm{W}$) (NVIDIA, 2023), the GPU-only energy is $1.96\,\mathrm{kW\,h}$; adding $40\,\mathrm{W}$ of non-GPU system draw yields $2.65\,\mathrm{kW\,h}$. Using the Netherlands gridmix factor 0.244 kg CO_2e/kWh (effective 1 January 2026) (CO2-emissiefactoren.nl, 2026), this corresponds to $0.48\,\mathrm{kg}$ CO_2e (GPU-only) or $0.65\,\mathrm{kg}$ CO_2e (including the system estimate), excluding embodied emissions.

8.3 ETHICAL IMPLICATIONS

The benchmark is synthetic and uses public datasets. Still, it models a real deployment risk: models can exploit non-causal correlates and fail under shift. In applied settings (for example, medical imaging), spurious correlates can encode confounders such as scanner type or hospital (Koh et al., 2021). Deep ensembles are attractive because they do not require explicit spurious-attribute labels, but their compute cost can limit accessibility.

8.4 CODE AND ARTIFACT AVAILABILITY

The code repository is publicly available at `https://github.com/jaouadhidayat/deep-ensembles-spurious-backgrounds`. The repository includes a release tagged `paper` that contains the code and artifacts used to produce the tables and figures in this paper.

Reproduction in three steps: (1) Generate the MNIST-on-CIFAR dataset instances using the provided dataset-generation script, (2) run the training sweeps for the specified (N, M) grids with the provided training entrypoint/configuration, and (3) generate the paper tables and figures using the provided artifact-generation script from the recorded runs.

8.5 LLM USE DISCLOSURE

We used a large language model as a writing and tooling assistant during drafting and revision. It was used to help turn feedback into a concrete edit list, to propose tighter wording for selected paragraphs, and to resolve presentation issues in LaTeX such as long URL wrapping in the references and keeping figures close to where they are first discussed. All experiments, plots, and analysis were produced from our code and recorded runs, and we only adopted suggestions that matched the evidence in the paper. Representative prompts are listed in Appendix B.

A IMPLEMENTATION DETAILS

Dataset generator: Each dataset instance is generated from a single random seed. We balance MNIST by subsampling each class to the minimum class count in the split (train and test are balanced separately). We then sample a pool of $10N$ CIFAR-10 images and reshape them into a $10 \times N$ class-specific "world" of backgrounds. Each class is assigned a disjoint set of N backgrounds.

For each MNIST example, we assign a CIFAR background index by sampling (with replacement) from the background set associated with its label. This assignment is fixed for the entire run. To build the composite input, we pad the MNIST digit to 32×32, convert it into a binary mask, and set the masked pixels in the CIFAR background to zero (a black silhouette).

OOD modes: **Seen-shuffle** draws OOD backgrounds from the same class-specific world but permutes the association between background sets and labels at test time. **Unseen-background** draws OOD backgrounds from a held-out CIFAR-10 pool disjoint from all backgrounds used in the class-specific world. This breaks the shortcut but also changes the background distribution relative to training.

Model and optimization: We train with SGD (momentum 0.9), learning rate 1.0, batch size 256, and 30 epochs. We use no dropout and no weight decay. Unless otherwise noted we use MSE against one-hot targets; we report cross-entropy as an ablation in the main paper.

Table 2: Training hyperparameters used in our runs.

Setting	Value
Optimizer	SGD
Learning rate	1.0
Momentum	0.9
Batch size	256
Epochs	30

Ensemble evaluation and subset variance: Within each run, we train the maximum ensemble size (LeNet: $M_{\max} = 8$; WRN-16-1: $M_{\max} = 2$) and evaluate smaller M as prefixes to keep the training protocol fixed. To check sensitivity to which members are chosen, we additionally evaluate 20 random subsets for intermediate M and compute the standard deviation of accuracy across subsets. Figure 7 reports this subset-variance diagnostic.

Confidence intervals: All reported 95% confidence intervals are computed with a hierarchical bootstrap that resamples dataset seeds (outer) and training seeds within each dataset (inner) with replacement. We use 2,000 bootstrap replicates.

Run coverage: We sweep $N \in \{1, 2, 4, 8, 16, 32, 64\}$. For LeNet, each (N, M) point averages 5 dataset seeds and 4 training initializations per dataset (20 runs). For WRN-16-1 and WRN-28-1, each point averages 5 dataset seeds and 2 training initializations per dataset (10 runs).

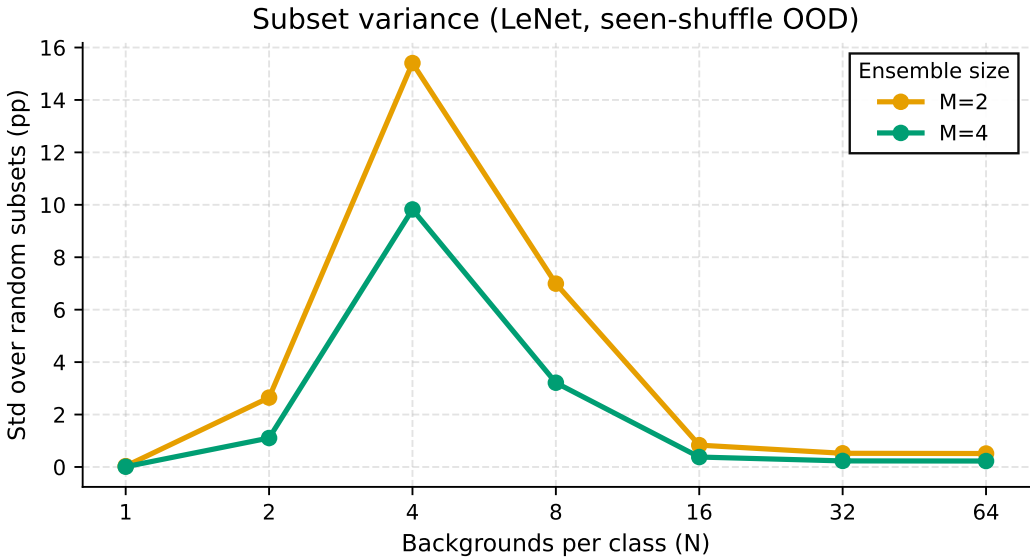

Figure 7: Subset variance diagnostic for LeNet under seen-shuffle OOD. Each point shows the standard deviation (in percentage points) of OOD accuracy across 20 random subsets of size M sampled from an $M{=}8$ ensemble. Sensitivity to subset choice is highest at intermediate N, where ensemble gains are also largest.

WRN filtering for stability: For WRN models trained with MSE, a subset of runs does not fit the correlated test distribution (unusually low correlated-test accuracy). To keep architecture comparisons interpretable, we exclude rows with correlated-test accuracy below 0.90 and report excluded counts per N.

Table 3: Excluded WRN rows due to low correlated-test accuracy (threshold < 0.90).

N	Model	Excluded	Total	Kept
1	WRN-16-1	0	20	20
2	WRN-16-1	0	20	20
4	WRN-16-1	0	20	20
8	WRN-16-1	0	20	20
16	WRN-16-1	6	20	14
32	WRN-16-1	9	20	11
64	WRN-16-1	10	20	10
1	WRN-28-1	0	10	10
2	WRN-28-1	0	10	10
4	WRN-28-1	0	10	10
8	WRN-28-1	1	10	9
16	WRN-28-1	1	10	9
32	WRN-28-1	1	10	9
64	WRN-28-1	1	10	9

Ensemble gain summary: Figure 8 reports the ensemble gain in OOD accuracy relative to a single model, as a function of N. This view highlights (i) that gains peak in an intermediate-N regime and can be negative at very small N, and (ii) that the location and magnitude of the peak depend on the loss (MSE versus CE) and on model family.

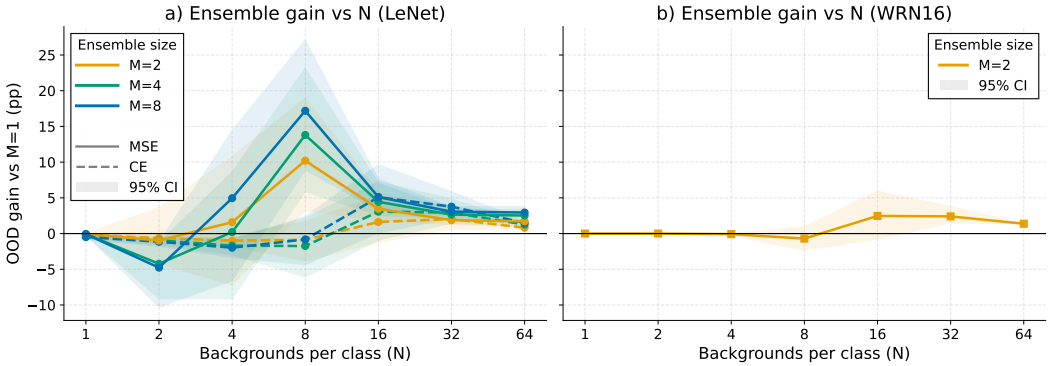

Figure 8: Ensemble gain versus N (seen-shuffle OOD). Left: LeNet gains for MSE (solid) and CE (dashed), shown for $M \in \{2, 4, 8\}$. Right: WRN-16-1 gains for $M{=}2$ (MSE). Shaded regions are 95% bootstrap confidence intervals.

B LLM Prompt Log

We used a large language model for targeted writing and tooling tasks. The prompts below are representative examples from the drafting process.

- "Turn these feedback notes into a short checklist of edits grouped by section."
- "Rewrite this paragraph to be clearer and shorter. Keep the meaning and keep all numbers unchanged."
- "Suggest two ways to explain the main trend in the results and one simple check for each."
- "Check this methodology text for missing details that a reader would need to reproduce the experiment."
- "Suggest a one sentence caption for this figure that explains what is plotted and what to notice."
- "Suggest a LaTeX fix so long URLs in the references do not overflow the margins."
- "Suggest LaTeX settings to place figures closer to their first mention without changing the content."
- "Scan this section for claims that need a citation or a pointer to a figure or table."

References

Taiga Abe, E. Kelly Buchanan, Geoff Pleiss, and John P. Cunningham. Deep ensembles work, but are they necessary? In *Advances in Neural Information Processing Systems*, 2022. URL https://papers.neurips.cc/paper_files/paper/2022/hash/da18c47118a2d09926346f33bebde9f4-Abstract-Conference.html.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. URL https://arxiv.org/abs/1907.02893.

CO2-emissiefactoren.nl. Elektriciteit: Stroom (Onbekend) Gridmix, 2026. URL https://co2emissiefactoren.nl/factoren/2026/11/52/elektriciteit-stroom-onbekend-gridmix/.

Elliot Creager, Joern-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2189–2200, 2021. URL https://proceedings.mlr.press/v139/creager21a.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43, 2020. URL https://jmlr.org/papers/v21/20-312.html.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017. URL https://arxiv.org/abs/1704.00109.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018. URL https://arxiv.org/abs/1803.05407.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/cifar.html.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826, 2021. URL https://proceedings.mlr.press/v139/krueger21a.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017. URL https://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019. doi: 10.1038/s41467-019-08987-4. URL https://doi.org/10.1038/s41467-019-08987-4.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://doi.org/10.1109/5.726791.

Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, 2021. URL https://arxiv.org/abs/2107.09044.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. *arXiv preprint arXiv:2007.02561*, 2020. URL https://arxiv.org/abs/2007.02561.

NVIDIA. GeForce RTX 4070 Laptop GPU specifications, 2023. URL https://www.nvidia.com/en-us/geforce/laptops/compare/.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/pdf?id=ryxGuJrFvS.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020. URL https://arxiv.org/abs/2006.09994.

Wenqian Ye, Luyang Jiang, Eric Xie, Guangtao Zheng, Yunsheng Ma, Xu Cao, Dongliang Guo, Daiqing Qi, Zeyu He, Yijun Tian, Megan Coffee, Zhe Zeng, Sheng Li, Ting-hao Huang, Ziran Wang, James M. Rehg, Henry Kautz, and Aidong Zhang. The clever hans mirage: A comprehensive survey on spurious correlations in machine learning. *arXiv preprint arXiv:2402.12715*, 2024. URL https://arxiv.org/abs/2402.12715.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. URL https://arxiv.org/abs/1605.07146.