Robust Finetuning from Generative Representations via Nuisance-extended Information Bottleneck

Minsun Lee, Gwangyeol Yu, Yunhun Nam, Jinwoo Shin, and Jongheon Jeong

***Abstract*—Models trained on limited or biased data can over-rely on target-correlated shortcuts or nuisance signals, leading to non-robust inferences under distribution shifts. Not only models trained from scratch, but also pretrained visual representations such as DINO, can retain shortcut factors during downstream fine-tuning. In this paper, we introduce an extension of the information bottleneck (IB) principle to address the shortcut problem, coined *nuisance-extended information bottleneck* (NIB). Specifically, NIB reformulates the compression process of the IB as a competition between semantic and nuisance bottlenecks. This competition encourages task-relevant information to remain in the semantic bottleneck while isolating nuisance variations, leading to more generalizable representations. We present two instantiations of NIB: (a) a vector quantization based nuisance bottleneck for training from scratch, and (b) a linear modeling upon the recent paradigm of representation autoencoder suitable for fine-tuning from modern pretrained representations. Our experiments show that NIB consistently improves the robustness and generalization of both from-scratch and pretrained representations across diverse downstream settings, notably without relying on task-specific robustness priors such as data augmentation. For example, NIB-based LoRA fine-tuning on ImageNet improves the average OOD performance of DINOv2 and SigLIP2 by 26.2% and 31.2%, respectively, compared to the standard LoRA adaptation.**

***Index Terms*—Information Bottleneck, Robust Fine-Tuning, Disentangled Representation Learning, Generative Representations.**

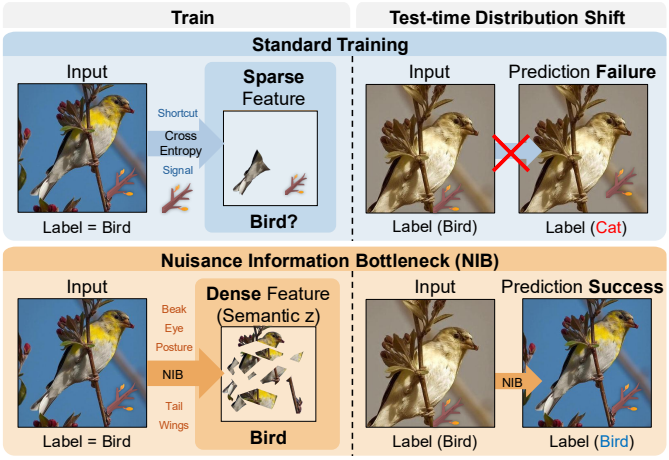


Fig. 1. Illustration of the *nuisance-extended information bottleneck* (NIB). We consider a test-time scenario where an input image may be corrupted while preserving its semantic content. In contrast to standard training, NIB aims to capture *all* target-correlated signals in the input image, including those that may remain reliable under distribution shifts.

I. Introduction

DEEP learning has driven rapid progress across a broad range of computer vision tasks, including image and video recognition [1]–[3], image synthesis [4]–[9], and 3D scene representation, rendering [10]–[14] and video generation [15], [16]. Despite these advances, deploying deep learning models in the real-world remains challenging because their predictions can be fragile under distribution shifts. In particular, when training data are limited or biased, models may make unreliable predictions for out-of-distribution (OOD) inputs even when those inputs are semantically close to in-distribution samples for humans [17], [18]. This issue is particularly critical in risk-sensitive applications such as autonomous driving [19], [20], medical imaging [21], [22], and healthcare [23]. These observations suggest that models often exploit shortcut signals rather than relying on semantic evidence, for example, latching onto background or texture cues instead of object identity [24].

Foundation visual representations have brought a breakthrough on this issue. For instance, CLIP [25] has shown that scaling up image-text contrastive learning can yield visual representations with unprecedented levels of robustness. Subsequent works on large-scale vision encoders, *e.g.*, the DINO [26]–[28] and SigLIP [29], [30] families, have further advanced the quality of pretrained visual representations in terms of their performance when applied to diverse downstream tasks. As a result, computer vision problems in the current paradigm are now largely approached through a simple adaptation of such foundation vision models, rather than training task-specific visual encoders from scratch [31].

However, adaptation to downstream tasks, especially under limited or biased supervision, can amplify shortcut factors already present in pretrained representations, thereby undermining their inherited robustness. Existing robust fine-tuning methods [32], [33] primarily focus on addressing this issue in zero-shot vision language models (*e.g.*, CLIP [25]), *e.g.*, by preserving the prior driven from the zero-shot model from image-text correspondence. Yet, this leaves open the question

Minsun Lee, Gwangyeol Yu, Yunhun Nam and Jongheon Jeong are with the Department of Artificial Intelligence, Korea University, Seoul, South Korea (E-mail: minsun3054@korea.ac.kr; rhkdduf627@korea.ac.kr; yh0326@korea.ac.kr; jonghj@korea.ac.kr).

Jinwoo Shin is with the Kim Jaechul Graduate School of Artificial Intelligence, Korea Advanced Institute of Science and Technology, Seoul, South Korea, and the School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea (E-mail: jinwoos@kaist.ac.kr).

This work was supported by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by Defense Acquisition Program Administration (DAPA) and Agency for Defense Development (ADD) (UD230017TD, 30%), Institute of Information & communications Technology Planning & Evaluation (IITP) grant (IITP-2026-RS-2025-02304828, 25%; RS-2019-II190079, 10%) and National Research Foundation of Korea (NRF) grant (RS-2025-23523603, 25%) funded by the Korea government (MSIT), and Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00345025, 10%) (Corresponding author: Jongheon Jeong).

of how to robustly adapt general pretrained vision encoders, *e.g.*, the DINO families [26]–[28], whose representations are not necessarily tied to zero-shot text alignment. Moreover, in domains where foundation models are difficult to apply, it remains an open challenge to learn robust representations without relying on domain-specific priors such as data augmentation.

A. Contributions

In this paper, we introduce the *Nuisance-extended Information Bottleneck* (NIB) as a scalable framework for robust fine-tuning of pretrained visual representations against unknown distribution shifts. The original work that proposed NIB [34] extends the classical Information Bottleneck (IB) principle [35] to the setting of adversarial input corruptions by decomposing the representation into two complementary bottlenecks: a semantic bottleneck for the downstream task and a nuisance bottleneck for residual variations. In this way, the semantic bottleneck effectively prevents excessive reliance on a subset of predictive signals, which would otherwise lead to shortcut overfitting, as shown in Figure 1.

While the underlying principle of NIB is broadly applicable, its practical instantiation was constrained by the representation learning paradigm available at the time. In particular, NIB requires a latent space that can simultaneously preserve semantic information for downstream prediction and nuisance information for reconstruction. Consequently, the original implementation was developed under an end-to-end representation learning paradigm, where latent representations were jointly learned and nuisance information could be explicitly preserved through pixel-space reconstruction. Extending the framework to modern pretrained vision models remained challenging because their representations were primarily optimized for discriminative tasks rather than reconstruction.

Recent advances in Representation Autoencoders (RAEs) [36] show that pretrained visual representations are sufficiently expressive to support faithful reconstruction. This finding removes a key obstacle to applying NIB to pretrained vision models, as pretrained representations can now serve as latent spaces that preserve both semantic and nuisance information. Motivated by this advance, we extend NIB to modern pretrained vision models. Specifically, we show that, together with RAEs, pretrained visual representations such as the DINO and SigLIP families admit an efficient realization of NIB, requiring only lightweight architectural adaptations to model semantic and nuisance representations.

We evaluate the effectiveness of NIB across a broad range of reliability benchmarks, covering both scenarios of fine-tuning pretrained vision encoders as well as training from scratch. Across domain generalization benchmarks, NIB improves the average OOD accuracy of DINOv2-B/14 [27] by 5.7% in the frozen-feature setting and by 26.2% under LoRA adaptation. For SigLIP2-B/16 [30], NIB improves the average OOD accuracy by 11.3% in the frozen-feature setting and by 31.2% under LoRA adaptation. This demonstrates that NIB is broadly applicable across various models. On fine-grained recognition benchmarks, NIB further improves the average transfer accuracy by 2.1% for DINOv2-B/14 and by 2.5% for SigLIP2-B/16, indicating that the semantic latent preserves class-discriminative information required for downstream recognition. The same principle is also effective for learning representations from scratch, reducing the CIFAR-10-C error by 23.1% and improving AUROC on the OBJECTS benchmark by up to 22.1%, without relying on additional task-specific robustness priors such as data augmentation.

B. Scope and Extensions

This paper extends a previous work appeared in the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition [34]. Compared to the conference publication, the main extensions include:

- A new implementation of the NIB framework for modern foundation models, enabled by recent advances in RAEs, replacing the original convolution-based realization.
- Substantially expanded and modernized experimental evaluation, reflecting current pretrained representation learning settings with foundation models (*e.g.*, DINO and SigLIP), additional datasets, adaptation scenarios, and robustness benchmarks.
- New ablation studies, analyses, and discussions, providing deeper insights into NIB and its design choices.
- A thoroughly revised manuscript, including reorganized technical presentation, redesigned figures, updated methodology and implementation details, and revised experimental sections centered on modern adaptation scenarios.

Overall, a substantial portion of the manuscript consists of new material that was not included in the conference publication. The journal version significantly extends the original work in terms of methodology, implementation, experimental validation, and technical presentation.

II. Related work

A. Robust Representation Learning

The reliability of visual representations has been extensively studied in various aspects of out-of-distribution (OOD) setups. For example, corruption robustness evaluates model stability under common corruptions such as noise, blur, fog, and changes in brightness [18], [37]–[39], while adversarial robustness instead considers worst-case input variations within a prescribed perturbation budget [40]–[42]. OOD detection aims to identify inputs that do not follow the training distribution [43]–[46]. Other lines of work study robustness to background, style, acquisition, or domain shifts, which are often closer to the distribution changes encountered in real deployment [38], [47]–[51]. Recent benchmarks have further broadened the scope of reliability evaluation from synthetic corruptions and isolated OOD datasets to real-world shifts, large-scale datasets, and foundation-model settings [49], [50]. Together with prior robustness studies, these benchmarks reveal a common vulnerability across different types of shifts: neural networks can rely on target-correlated shortcut signals or nuisance-sensitive cues whose validity weakens under distribution shifts [38], [47]–[50], [52], [53]. Despite this progress, many robustness methods are designed around specific assumptions about the shift. For

example, corruption-robust methods often rely on predefined corruption types or augmentation strategies [37], [54], OOD detection methods may depend on particular scoring rules or auxiliary outlier data [43]–[45], and adversarially robust models are usually optimized for a specific threat model [40]. Rather than addressing each shift type with a separate assumption, we address this problem by leveraging the information bottleneck principle, which encourages representations to retain information relevant to the task, and extending it to explicitly separate semantics relevant to the task from nuisance information.

B. Generative Visual Representations

Visual representation learning now commonly leverages foundation visual representations for adaptation to downstream tasks, rather than training task-specific encoders from scratch. The DINO and SigLIP families provide transferable visual features for classification, retrieval, and dense recognition [26], [27], [29], [30]. In addition to image-level embeddings, modern pretrained visual encoders also provide structured token-level features, which are useful for spatially grounded downstream tasks [27], [30], [55], [56]. This trend also extends to generative modeling, where REPA and RAE use strong pretrained visual representations as generative latent [36], [57]. These advances suggest that modern pretrained representations can serve both discriminative and generative purposes. However, using all encoded information during adaptation can encourage shortcut reliance, as cues predictive in the in-distribution may become unreliable under distribution shifts. Thus, a key challenge is to separate prediction-relevant semantics from residual variation that should be preserved but not used for prediction, thereby improving downstream reliability under distribution shifts.

C. Robust Adaptation of Pretrained Representations

Adapting pretrained models to downstream tasks can improve in-distribution accuracy, but it may degrade OOD generalization [58], [59]. Robust fine-tuning studies how to adapt pretrained models while preserving their generalization under distribution shifts. WiSE-FT interpolates between zero-shot and fine-tuned weights to balance downstream accuracy and OOD robustness [32], while FLYP retains a pretraining-style contrastive objective during fine-tuning [33]. Surgical fine-tuning shows that the choice of which layers to adapt can have a significant effect under distribution shifts [60]. More recent approaches further introduce text-guided regularization or parameter-efficient adaptation for robust fine-tuning [61], [62]. These works provide important tools for adapting pretrained models, but many of them are centered on vision-language models and rely on zero-shot priors, text-aligned representations, or multimodal alignment [25], [32], [33], [61], [62]. In contrast, pretrained vision encoders such as DINO cannot rely on the text-based priors used in CLIP-based robust fine-tuning. Even text-aligned encoders may not always have access to suitable external prompts. Accordingly, there remains a need for a more generally applicable adaptation scheme that directly improves the reliability of pretrained visual representations.

D. Hybrid Discriminative-Generative Modeling

Another related direction combines discriminative prediction with generative modeling. Prior work has incorporated generative models into classifiers by placing density models on learned representations, or by interpreting classifiers as energy-based models that can model both prediction and data likelihood [63]–[65]. However, generative modeling alone does not guarantee semantically reliable representations. Likelihood-based models can assign high likelihood to OOD samples and can be strongly affected by low-level background or texture statistics [66]–[69]. Joint energy-based models also demonstrate the potential of combining discriminative and generative objectives, but their training can be unstable and difficult to scale [63], [70]. Unlike such hybrid discriminative-generative models, the proposed approach explicitly separates prediction-relevant semantics from residual information preserved for reconstruction.

E. Information Bottleneck Based Modeling

The information bottleneck (IB) principle was introduced as a way to learn representations that compress the input while retaining information needed to predict a target [35], [71]. Variational information bottleneck (VIB) made this principle practical for deep neural networks through variational approximation and KL regularization [72]. The IB perspective has since been used to study generalization, robustness, disentanglement, and domain generalization. Disentangled Information Bottleneck (DisenIB), for example, factorizes representations into task-relevant and task-irrelevant components [73]. Other approaches, including IRM, DANN, Learning Not to Learn, and invariant information bottleneck methods, aim to suppress domain-specific or bias-sensitive cues and learn invariant representations under distribution shifts [74]–[79]. Unlike many prior IB-based approaches that remove or suppress task-irrelevant or domain-specific cues, the proposed approach separates such residual information from the predictive representation rather than discarding it. This enables nuisance information to be preserved without interfering with prediction, while maintaining a semantically focused representation for downstream tasks.

III. Preliminaries

A. Notations

Given two random variables $x \in \mathcal{X}$, the input, and $y \in \mathcal{Y}$, the target, our goal is to learn a mapping, or an encoder, $f : \mathcal{X} \to \mathcal{Z}$ from data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ such that the representation $z := f(x)$ can predict y through a simpler, *e.g.*, linear, mapping [80], [81]. We assume that f is parametrized by a neural network and define it stochastically to adopt an information-theoretic view [71]. Namely, the encoder output is treated as a random variable defined by $p_f(z|x)$ rather than as a deterministic constant. This stochastic modeling can be implemented through the reparameterization trick [82] with an independent random variable ϵ and a deterministic mapping f by assuming $z := f(x, \epsilon)$. For example, a standard design parametrizes f as

$$f(\mathbf{x}, \boldsymbol{\epsilon}) := f^{\mu}(\mathbf{x}) + \boldsymbol{\epsilon} \cdot f^{\sigma}(\mathbf{x}), \tag{1}$$

where $f^{\mu} \in \mathbb{R}^{|\mathcal{Z}|}$ and $f^{\sigma} \in \mathbb{R}_{+}^{|\mathcal{Z}|}$ are deterministic mappings that model μ and σ in $\mathcal{N}(x; \mu, \sigma^2 I)$, respectively, so that they can be learned through gradient-based optimization.

The data $\mathcal{D}$ is commonly assumed to consist of i.i.d. samples from a data-generating distribution, *i.e.*, $(x_i, y_i) \sim p_d(x, y)$. One expects that an encoder f learned from $\mathcal{D}$ can generalize to predict $p_d(y|x)$ for unseen samples from $p_d(x, y)$. However, this formulation does not specify how f should behave for inputs that are unlikely to be drawn from p_d, denoted by $\hat{x}$. This becomes problematic when one expects the decision making of f to be close to that of humans, at least when $\hat{x}$ differs from the training distribution only by changes that humans regard as nuisance. This is where current neural networks often fail under standard training practices.

B. Information Bottleneck

Intuitively, a good representation z should preserve the information in x that is correlated with y, while preventing z from becoming overly complex. The Information Bottleneck (IB) principle [35], [71] provides a principled way to obtain such a succinct representation $x \to z$ for a given downstream task $x \to y$. Specifically, it seeks a representation z that maximizes the task-relevant mutual information $I(z; y)$ while minimizing $I(x; z)$ to constrain the capacity of z for better generalization. In this sense, the mutual information $I(x; z)$ serves as a complexity measure of z. The standard IB objective is written as

$$\max_{f} R_{\texttt{IB}}(f), \quad \text{for} \quad R_{\texttt{IB}}(f) := I(\mathbf{z}; \mathbf{y}) - \beta I(\mathbf{x}; \mathbf{z}), \tag{2}$$

where $\beta \geq 0$ controls the capacity constraint, ensuring $I(x; z) \leq I_\beta$ for some I_β.

Computing the mutual information of two random variables is generally hard and this makes the IB objective infeasible in practice. To overcome this, *variational information bottleneck* (VIB) [72], [83] applies variational inference to obtain a lower bound on the IB objective (2). Specifically, it approximates: (a) $p(\mathbf{y}|\mathbf{z})$ by a (parametrized) "decoder" neural network $q(\mathbf{y}|\mathbf{z})$, and (b) $p(\mathbf{z})$ by an "easier" distribution $r(\mathbf{z})$, *e.g.*, isotropic Gaussian $\mathcal{N}(\mathbf{z}|0, \mathbf{I})$. Having such (variational) approximations in computing (2) as well as the Markov chain property $\mathbf{y} - \mathbf{x} - \mathbf{z}$ of neural networks, one yields the following lower bound on the IB objective (2):

$$\begin{aligned} &I(\mathbf{z}; \mathbf{y}) - \beta I(\mathbf{z}; \mathbf{x}) \\ &\quad \geq \mathbb{E}_{\mathbf{x},\mathbf{y}} \left[\int dz \left(p(z|\mathbf{x}) \log q(\mathbf{y}|z) - \beta p(z|\mathbf{x}) \log \tfrac{p(z|\mathbf{x})}{r(z)} \right) \right]. \end{aligned} \tag{3}$$

This bound can be approximated with the empirical distribution $p(x, y) \approx \frac{1}{n} \sum_i \delta_{x_i}(x) \delta_{y_i}(y)$ from data. By further assuming a Gaussian encoder $p(z|x) := \mathcal{N}(z|f^{\mu}(x), f^{\sigma}(x))$ as defined above and applying the reparameterization trick [82], we obtain the following VIB objective:

$$L^{\beta}_{\texttt{VIB}} := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\epsilon}}[-\log q(y_i|f(x_i, \boldsymbol{\epsilon}))] + \beta \, \mathrm{KL}\left(p(\mathbf{z}|x_i) \| r(\mathbf{z})\right). \tag{4}$$

This objective provides a practical surrogate for the original IB objective. The first term encourages the representation z to predict the target y, while the second term regularizes the information capacity of z through a KL penalty.

IV. Method

The standard Information Bottleneck (IB) objective learns a representation $\mathbf{z} := f(\mathbf{x})$ under the premise that future inputs are also drawn from the same data-generating distribution $p_d(\mathbf{x}, \mathbf{y})$. In practice, however, a test-time input may not be observed exactly as in the training distribution. Instead, an input $\mathbf{x}$ can be transformed into $\hat{\mathbf{x}}$ through an unknown noisy channel. We assume that this transformation preserves the semantic information relevant to the target $\mathbf{y}$, namely,

$$I(\mathbf{x}; \mathbf{y}) = I(\hat{\mathbf{x}}; \mathbf{y}) > 0. \tag{5}$$

This setting can be understood as a case where $\mathbf{x}$ contains multiple signals that are individually correlated with $\mathbf{y}$. Some of these signals may be removed without changing the mutual information between the input and the target. Such target-correlated signals are commonly exploited by deep neural networks: signals that appear as nuisances to humans can still provide sufficient information for prediction. Therefore, a robust representation should not rely on a single shortcut-like cue, but should instead capture target-relevant information in a stable and comprehensive manner.

From the perspective of IB, if the goal is to learn a succinct encoder f, the unknown noisy channel can be viewed as an adversary. That is, one may consider a transformed input $\hat{\mathbf{x}}$ that preserves the semantic information about $\mathbf{y}$ while minimizing the target information retained in the learned representation. This can be written as the following adversarial formulation:

$$\min_{\hat{\mathbf{x}}} I(\hat{\mathbf{z}} := f(\hat{\mathbf{x}}); \mathbf{y}) \;\; \text{subject to} \;\; I(\mathbf{x}; \mathbf{y}) = I(\hat{\mathbf{x}}; \mathbf{y}), \tag{6}$$

This formulation assumes that we do not have prior knowledge of how the noisy channel behaves. Under this assumption, the encoder f is encouraged to extract all signals in $\mathbf{x}$ that are highly correlated with $\mathbf{y}$. Otherwise, the adversarial noisy channel could remove all signals except the one that f fails to capture, causing $\hat{\mathbf{z}} = f(\hat{\mathbf{x}})$ to lose information necessary for target prediction. Directly optimizing this adversarial objective, however, is computationally infeasible, as it involves an almost unconstrained optimization over the high-dimensional input space $\mathcal{X}$, with an intractable mutual information constraint.

A. Nuisance-extended Information Bottleneck

To encourage such adversarial behavior without directly solving the above problem, we explicitly model a nuisance representation $\mathbf{z}_n$ in addition to the semantic representation $\mathbf{z}$. The nuisance representation $\mathbf{z}_n$ is intended to encode the residual information that is not contained in $\mathbf{z}$ but is still needed to reconstruct $\mathbf{x}$. Thus, the pair $(\mathbf{z}, \mathbf{z}_n)$ should preserve input information, which corresponds to maximizing $I(\mathbf{x}; \mathbf{z}, \mathbf{z}_n)$. At the same time, $\mathbf{z}_n$ should not contain discriminative information about $\mathbf{y}$, which corresponds to minimizing $I(\mathbf{z}_n; \mathbf{y})$.

Consequently, information correlated with $\mathbf{y}$ should be encoded complementarily in the semantic representation $\mathbf{z}$,

rather than remaining in the nuisance branch. In this view, the capacity constraint in the original IB objective plays a more central role: it not only regularizes $\mathbf{z}$ to be compact, but also controls the allocation of information between $\mathbf{z}$ and $\mathbf{z}_n$. This creates a competition between the two information channels, where $\mathbf{z}_n$ preserves residual information for reconstruction while $\mathbf{z}$ retains the information necessary for prediction.

We therefore define the *nuisance-extended information bottleneck* (NIB) objective by combining the standard IB objective with nuisance modeling:

$$\max_f R_{\texttt{NIB}}(f) := R_{\texttt{IB}}(f) - I(\mathbf{z}_n;\mathbf{y}) + \alpha I(\mathbf{x};\mathbf{z},\mathbf{z}_n), \quad (7)$$

where $\alpha \geq 0$. The NIB objective can be viewed as a regularized extension of IB with an explicit nuisance representation. In particular, it encourages maximizing $I(\mathbf{x};\mathbf{z},\mathbf{z}_n)$ while minimizing $I(\mathbf{z}_n;\mathbf{y})$. In the ideal case, these correspond to $H(\mathbf{x}|\mathbf{z},\mathbf{z}_n) = 0$ and $I(\mathbf{z}_n;\mathbf{y}) = 0$, respectively. In other words, $\mathbf{z}$ and $\mathbf{z}_n$ should jointly preserve the input, while $\mathbf{z}_n$ itself should not be informative about the label.

The following lemma formalizes why these conditions, together with independence between the semantic and nuisance representations, are sufficient for recovering the original target-relevant information from the semantic representation.

Lemma 1. *Let $\mathbf{x} \in \mathcal{X}$, and $\mathbf{y} \in \mathcal{Y}$ be random variables, $\hat{\mathbf{x}}$ be a noisy observation of $\mathbf{x}$ with $I(\mathbf{x};\mathbf{y}) = I(\hat{\mathbf{x}};\mathbf{y})$. Given that a representation $[\hat{\mathbf{z}},\hat{\mathbf{z}}_n] := f(\hat{\mathbf{x}})$ of $\hat{\mathbf{x}}$ satisfies (a) $H(\hat{\mathbf{x}}|\hat{\mathbf{z}},\hat{\mathbf{z}}_n) = 0$, (b) $I(\hat{\mathbf{z}}_n;\mathbf{y}) = 0$, and (c) $\hat{\mathbf{z}} \perp \hat{\mathbf{z}}_n$, it holds $I(\hat{\mathbf{z}};\mathbf{y}) = I(\mathbf{x};\mathbf{y})$.*

The proof of Lemma 1 is provided in the supplementary material. Lemma 1 indicates that a robust semantic representation requires three properties: reconstruction, nuisance suppression, and independence. The semantic and nuisance latents should jointly preserve the input, the nuisance latent should remove label-discriminative information, and the two latents should remain independent. Accordingly, $\mathbf{z}_n$ absorbs residual nuisance information required for reconstruction, while $\mathbf{z}$ preserves the semantic information required for target prediction.

B. Practical Autoencoder-Based Objective

Based on the NIB objective and Lemma 1, we instantiate NIB as a practical variational autoencoder-based framework, which we call *nuisance-extended variational IB* (NVIB). Specifically, NVIB considers an encoder $f : \mathcal{X} \to \mathcal{Z}$ together with a decoder $g : \mathcal{Z} \to \mathcal{X}$, and extends VIB by implementing the three desiderata suggested by Lemma 1 through trainable objectives: (a) joint reconstruction by $\mathbf{z}$ and $\mathbf{z}_n$, (b) nuisance-ness of $\mathbf{z}_n$ with respect to $\mathbf{y}$, and (c) independence between $\mathbf{z}$ and $\mathbf{z}_n$.

Firstly, we adopt a reconstruction loss $L_{\texttt{recon}}$ to encourage $\mathbf{z}$ and $\mathbf{z}_n$ to jointly reconstruct $\mathbf{x}$, thereby maximizing $\log p(\mathbf{x}|\mathbf{z},\mathbf{z}_n)$. Under the standard assumption that the decoder output follows $\mathcal{N}(\mathbf{x},\sigma^2\mathbf{I})$, this corresponds to a mean-squared error objective. This term implements the $I(\mathbf{x};\mathbf{z},\mathbf{z}_n)$ component of NIB, encouraging nuisance information to be preserved in $\mathbf{z}_n$ rather than simply discarded.

Second, to enforce the nuisance property of $\mathbf{z}_n$ with respect to $\mathbf{y}$, we approximate $p(\mathbf{y}|\mathbf{z}_n)$ with a multi-layer perceptron (MLP) q_n and train it adversarially. The classifier q_n is first optimized to predict $\mathbf{y}$ from $\mathbf{z}_n$, while the encoder is trained so that the prediction from $\mathbf{z}_n$ becomes close to the uniform distribution over $\mathcal{Y}$. This can be formulated by the following:

$$L_{\texttt{nuis}} := \mathbb{E}_{\mathbf{x}}[\mathbb{CE}(q_n^*(\mathbf{z}_n), \tfrac{\mathbf{1}}{|\mathcal{Y}|})], \quad \text{where } q_n^* := \arg\min_{q_n} \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbb{CE}(q_n(\mathbf{z}_n),\mathbf{y})], \quad (8)$$

where $\mathbb{CE}$ denotes the cross-entropy loss. This objective lowers $I(\mathbf{z}_n;\mathbf{y})$ by making $\mathbf{z}_n$ useful for reconstruction but uninformative for label prediction.

Third, to encourage independence between $\mathbf{z}$ and $\mathbf{z}_n$, we assume that their joint prior follows an isotropic Gaussian distribution, *i.e.*, $p(\mathbf{z},\mathbf{z}_n) \sim \mathcal{N}(0,\mathbf{I})$, and apply either a GAN-based loss (Section IV-C1) or a SIGReg-based loss [84] (Section IV-C2), both denoted by $L_{\texttt{ind}}$. These regularizers encourage $\mathbf{z}$ and $\mathbf{z}_n$ to form an approximately independent and tractable latent space, which is also useful for sampling from the learned decoder.

Lastly, to approximate the original IB term $R_{\texttt{IB}}(f)$ in the NVIB objective, we use the VIB objective $L^{\beta}_{\texttt{VIB}}$ [72], [83]. VIB approximates $p(\mathbf{y}|\mathbf{z})$ with a parameterized decoder network $q(\mathbf{y}|\mathbf{z})$, and approximates $p(\mathbf{z})$ with a tractable distribution $r(\mathbf{z})$ such as an isotropic Gaussian $\mathcal{N}(\mathbf{z}|0,\mathbf{I})$. Assuming a Gaussian encoder and applying the reparameterization trick, we obtain

$$L^{\beta}_{\texttt{VIB}} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\epsilon}}[-\log q(y_i|f(x_i,\boldsymbol{\epsilon}))] + \beta\ \mathrm{KL}\ (p(\mathbf{z}|x_i)\|r(\mathbf{z})). \quad (9)$$

This objective preserves the information in $\mathbf{z}$ that is useful for predicting $\mathbf{y}$, while regularizing the amount of input information encoded in $\mathbf{z}$.

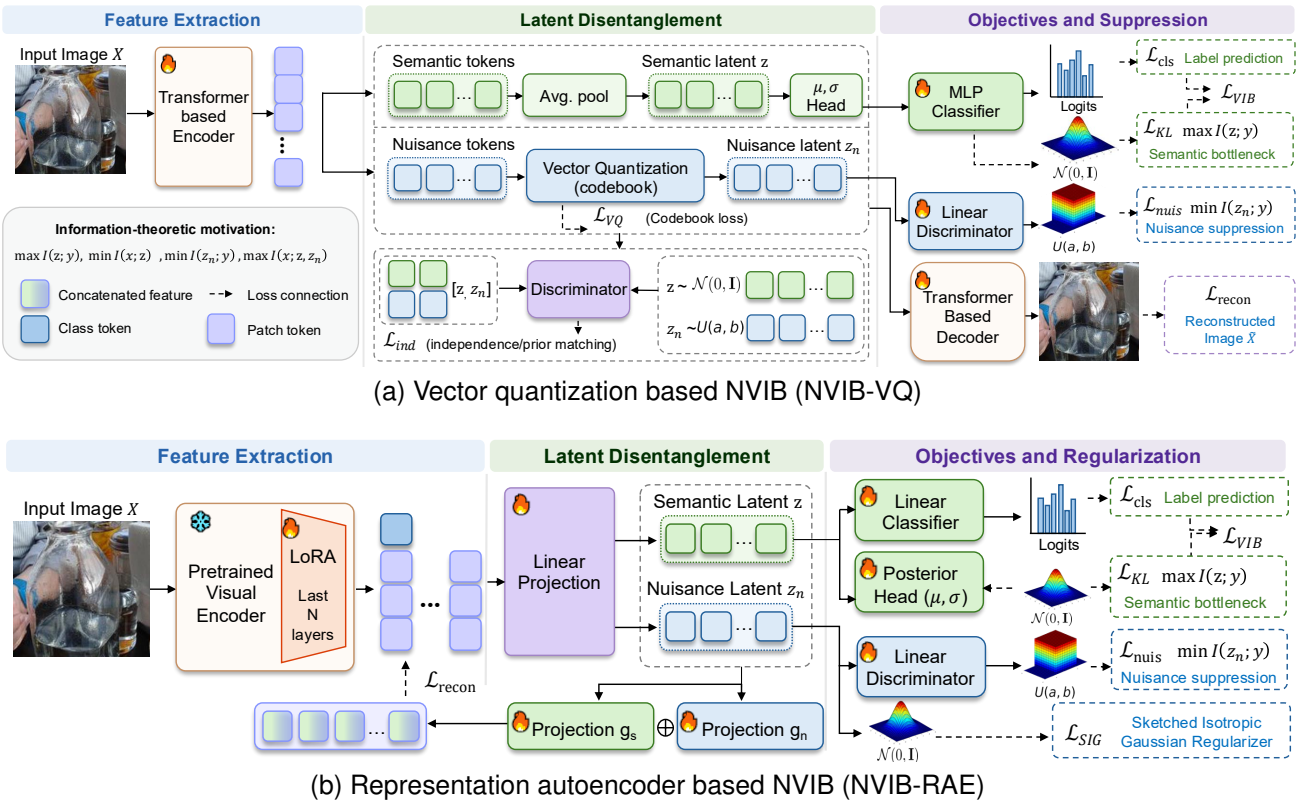
Combining these terms gives the final NVIB objective:

$$L_{\texttt{NVIB}} := L^{\beta}_{\texttt{VIB}} + \alpha \cdot L_{\texttt{recon}} + L_{\texttt{nuis}} + L_{\texttt{ind}}. \quad (10)$$

Here, α controls the relative strength of the reconstruction objective. This objective jointly optimizes semantic prediction, nuisance suppression, independence regularization, and input reconstruction in a single autoencoder-based framework.

C. Architectural Designs

We present two complementary architectures for implementing the NVIB objective (Section IV-B), depending on whether the model is trained from scratch or leverages pretrained representations (*e.g.*, DINO [26], [27] and SigLIP [29], [30]). For models trained from scratch, we employ a vector quantization (VQ)-based autoencoder design to model the nuisance representation, which stabilizes optimization during the early stages of training (Section IV-C1). For models leveraging pretrained representations, we draw inspiration from the recent success of Representation Autoencoders (RAEs) [36], which show that foundation representations are already sufficiently expressive for image reconstruction when coupled with an appropriate decoder. Building on this observation, we propose a linear-mapping based architecture to factorize pretrained representations into semantic and nuisance features (Section IV-C2). The following two subsections describe these architectures in detail.



(a) Vector quantization based NVIB (NVIB-VQ)

(b) Representation autoencoder based NVIB (NVIB-RAE)

Fig. 2. Overview of two NVIB instantiations, namely NVIB-VQ and NVIB-RAE. (a) NVIB-VQ uses a ViT-based image-token autoencoder with a vector-quantized nuisance branch. (b) NVIB-RAE applies the same semantic-nuisance decomposition to pretrained visual tokens. In both cases, the semantic latent $\mathbf{z}$ is used for prediction, while $\mathbf{z}$ and $\mathbf{z}_n$ jointly preserve information through the reconstruction objective.

*1) VQ-based Nuisance Modeling (**NVIB-VQ**):* NVIB-VQ instantiates NVIB with a Vision Transformer (ViT) [85] based image-token autoencoder whose key design is to discretize the patch-level nuisance branch using vector quantization, keeping the nuisance representation compact and its marginal distribution tractable while the semantic branch stays focused on target prediction. On top of this VQ-based branch, NVIB-VQ is trained with an SSIM-based reconstruction loss that preserves structural residual information and a GAN-based independence loss that decorrelates the semantic and nuisance latents. Figure 2 illustrates the resulting architecture, which separates image patch tokens into semantic and nuisance branches.

On the encoder side, the patch-token embedding is split into two branches with reduced embedding dimensions to model $\mathbf{z}$ and $\mathbf{z}_n$. One branch is average-pooled to define the semantic latent $\mathbf{z}$, while the other branch is vector-quantized to define the nuisance representation $\mathbf{z}_n$. Thus, the semantic branch summarizes global semantic information into a compact latent, whereas the nuisance branch preserves patch-level residual information. For the decoder side, on the other hand, linear projections map $\mathbf{z}$ and $\mathbf{z}_n$ back into the patch embedding space before they are processed by a Transformer decoder.

In this Transformer-based design, global average pooling is applied only to $\mathbf{z}$, making its dimensionality independent of the input resolution, while $\mathbf{z}_n$ remains at the patch level. As a result, the dimensionality of $\mathbf{z}_n$ increases with higher-resolution inputs, which can make training difficult in two ways. First, because NVIB-VQ essentially creates a competition between two information channels, $\mathbf{z}$ and $\mathbf{z}_n$, a large dimensional mismatch makes it harder to balance the objectives. Second, as $\mathbf{z}_n$ becomes higher-dimensional, it becomes increasingly difficult to regularize $\mathbf{z}_n$ toward an independent and tractable marginal distribution for sampling.

To alleviate these issues, we apply vector quantization (VQ) [86], [87] to the nuisance embedding. Let $\hat{\mathbf{z}}_{n,j}$ denote the pre-quantized nuisance embedding of the j-th patch token, and let $\mathbf{E} = \{\mathbf{e}_k\}_{k=1}^{K}$ be a learned codebook. Each nuisance embedding is quantized to its nearest code vector,

$$k_j^* = \arg\min_k \|\hat{\mathbf{z}}_{n,j} - \mathbf{e}_k\|_2^2, \qquad \mathbf{z}_{n,j} = \mathbf{e}_{k_j^*}. \tag{11}$$

This represents $\mathbf{z}_n$ as a sequence of discrete codes rather than continuous per-patch embeddings, enabling more scalable balancing with $\mathbf{z}$ and providing a tractable marginal distribution. Since $\mathbf{z}_n$ is discrete, post-hoc generative modeling, such as an autoregressive prior or a diffusion model, can be applied for efficient nuisance sampling.

Specifically, we add the following objective to the NVIB objective to enable VQ-based nuisance modeling:

$$\begin{aligned} L_{\mathtt{VQ}}(\mathbf{x}; \mathbf{E}) := \frac{1}{N} \sum_{j=1}^{N} & \left\| \mathrm{sg}[\hat{\mathbf{z}}_{n,j}] - \mathbf{e}_{k_j^*} \right\|_2^2 \\ & + \beta_{\mathtt{VQ}} \left\| \hat{\mathbf{z}}_{n,j} - \mathrm{sg}[\mathbf{e}_{k_j^*}] \right\|_2^2 . \end{aligned} \tag{12}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator defined by $\mathrm{sg}[x] \equiv x$ and $\frac{d}{dx}\,\mathrm{sg}[x] \equiv 0$.

Thus, NVIB-VQ can be summarized as optimizing the NVIB objective together with the VQ objective:

$$L_{\texttt{NVIB-VQ}} = L_{\texttt{NVIB}} + L_{\texttt{VQ}}. \tag{13}$$

In this architecture, $\mathbf{z}$ is used as a compact semantic latent for target prediction, while $\mathbf{z}_n$ is used as a discrete nuisance latent that preserves patch-level residual information for reconstruction and sampling. Consequently, Transformer-based NVIB-VQ separates semantic information from nuisance variation using token-level structure, while preventing residual information from being discarded through reconstruction.

a) SSIM-Based D_2^2 Reconstruction Loss: NVIB-VQ minimizes a reconstruction loss to implement the $I(\mathbf{x}; \mathbf{z}, \mathbf{z}_n)$ term in the NVIB objective. Although normalized mean-squared error (NMSE) is a natural default choice, the reconstruction loss is not limited to pixel-wise error. For Transformer-based NVIB-VQ, we use an SSIM-based reconstruction loss so that reconstruction can better reflect perceptual and structural similarity. In the supplementary material, we verify the effectiveness of this design in comparison to NMSE.

Let $\hat{\mathbf{x}} = g(\mathbf{z}, \mathbf{z}_n)$ denote the reconstructed image. For a given image and its reconstruction $(\mathbf{x}, \hat{\mathbf{x}})$, the structural similarity index measure (SSIM) defines a similarity metric by considering differences in luminance, represented by S_1, and structural consistency, represented by S_2:

$$\mathrm{SSIM}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{2\mu_{\mathbf{x}}\mu_{\hat{\mathbf{x}}} + c_1}{\mu_{\mathbf{x}}^2 + \mu_{\hat{\mathbf{x}}}^2 + c_1} \cdot \frac{2\sigma_{\mathbf{x}\hat{\mathbf{x}}} + c_2}{\sigma_{\mathbf{x}}^2 + \sigma_{\hat{\mathbf{x}}}^2 + c_2} =: S_1 \cdot S_2, \tag{14}$$

where $c_1 := 0.01^2$ and $c_2 := 0.03^2$ are small constants for numerical stability. In practice, SSIM can be computed on a per-patch basis using a sliding window and then averaged, and we follow this implementation.

Since SSIM itself is not a distance metric and can take negative values, we instead use the squared-D_2 distance, denoted by D_2^2, as the reconstruction loss:

$$D_2^2(\mathbf{x}, \hat{\mathbf{x}}) := \sqrt{(1 - S_1) + (1 - S_2)}^2 = 2 - S_1 - S_2. \tag{15}$$

This loss encourages structure-preserving reconstruction rather than relying only on pixel-wise error, thereby aligning the reconstruction branch with the goal of preserving residual information needed for semantic-nuisance separation.

b) GAN-based independence loss: To encourage independence between $\mathbf{z}$ and $\mathbf{z}_n$ in NVIB-VQ, we apply GAN-style prior matching to their joint latent distribution. Specifically, we match the encoded joint latent pair $(\mathbf{z}, \mathbf{z}_n)$ to a factorized isotropic Gaussian prior. Since the prior factorizes across the semantic and nuisance dimensions, matching the joint latent distribution to this prior encourages $\mathbf{z}$ and $\mathbf{z}_n$ to become approximately independent. Let q_{z} be a two-layer MLP discriminator that distinguishes encoded latent pairs from samples drawn from the isotropic Gaussian prior. For an encoded pair $(\mathbf{z}, \mathbf{z}_n) = f(\mathbf{x})$ and a prior sample $(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}_n) \sim \mathcal{N}(0, \mathbf{I})$, we define the GAN-based independence objective as

$$\begin{aligned} L_{\texttt{ind}} := \max_{q_z} \mathbb{E}_{\mathbf{x}} \left[\log q_z(\mathbf{z}, \mathbf{z}_n)\right] \\ + \mathbb{E}_{(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}_n) \sim \mathcal{N}(0, \mathbf{I})} \left[\log\left(1 - q_z(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}_n)\right)\right]. \end{aligned} \tag{16}$$

The discriminator is trained to distinguish the encoded latent pairs from Gaussian prior samples, while the encoder is trained adversarially so that $(\mathbf{z}, \mathbf{z}_n)$ becomes indistinguishable from the factorized prior. This regularizes the joint latent space toward an independent and tractable distribution, which supports both semantic-nuisance separation and latent sampling.

*2) RAE-based Nuisance Modeling (**NVIB-RAE**):* Pretrained visual representations have traditionally been viewed primarily as discriminative features for recognition, rather than as generative features that can support reconstruction. Therefore, generative representation learning has often relied on separately trained autoencoders to obtain reconstruction-compatible latent spaces. Recent Representation Autoencoders (RAEs) [36], however, show that sufficiently large pretrained visual representations, such as DINO-family features, can serve as expressive latent representations for reconstruction when coupled with an appropriate decoder.

This observation provides a simpler way to instantiate NVIB in the pretrained representation setting. Instead of learning a separate encoder solely to obtain a reconstruction-compatible latent space, NVIB-RAE performs semantic-nuisance factorization directly on pretrained visual tokens. As illustrated in Figure 2, NVIB-RAE separates pretrained token representations into semantic latents for downstream prediction and nuisance latents for preserving residual variation during training.

a) Latent-space Reconstruction: Unlike NVIB-VQ that is based on pixel-level reconstructions, NVIB-RAE learns a decomposition of semantic-nuisance features directly from reconstructing patch-level tokens. This is implemented as a set of simple linear heads on the token-wise representations. Following the VIB formulation in Eq. (9), a semantic linear branch predicts a mean $\mu_{i,j}$ and scale $\sigma_{i,j}$, from which $\mathbf{z}_{i,j}$ is sampled using the reparameterization trick. In contrast, the nuisance branch produces a deterministic nuisance code $\mathbf{z}_{n,i,j}$. In this way, nuisance-sensitive variations, such as background, local texture, illumination, and corruption, often manifest as spatial patterns or local token responses rather than as a single global signal.

Given a pretrained patch token $p_{i,j} \in \mathbb{R}^d$, the semantic and nuisance latents $\mathbf{z}_{i,j}$ and $\mathbf{z}_{n,i,j}$ are projected back to the patch-feature dimension by trainable reconstruction projections g_s and g_n. The reconstructed token is defined as:

$$\hat{p}_{i,j} = g_s(\mathbf{z}_{i,j}) + g_n(\mathbf{z}_{n,i,j}). \tag{17}$$

This additive reconstruction forces the semantic and nuisance branches to jointly preserve the pretrained representation, while allowing $\mathbf{z}_n$ to absorb reconstruction-relevant residual information.

For the reconstruction loss, we optimize the normalized MSE computed directly for the token representations:

$$L_{\texttt{recon}} = \frac{1}{\sigma_{\mathcal{B}}^2 + \epsilon_{\mathrm{num}}} \frac{1}{BNd} \sum_{i=1}^{B} \sum_{j=1}^{N} \|\hat{p}_{i,j} - p_{i,j}\|_2^2, \tag{18}$$

where $\sigma_{\mathcal{B}}^2$ denotes the empirical variance of the target pretrained patch features $\{p_{i,j}\}_{i=1,j=1}^{B,N}$ in the mini-batch, and $\epsilon_{\mathrm{num}} > 0$ is a small constant for numerical stability. This objective discourages residual information from being discarded and encourages $\mathbf{z}$ and $\mathbf{z}_n$ to jointly preserve the pretrained representation.

b) Patch-level Task Head: To implement the VIB term $L^{\beta}_{\texttt{VIB}}$ Eq. (9) for NVIB-RAE, we apply a classifier head $h_{\texttt{cls}}$ shared across the semantic patch latents $\mathbf{z}_{i,j}$. The classifier produces a patch-level class prediction, and the image-level prediction is obtained by averaging the patch-level predictions across all patch tokens.

$$q_{i,j} = \mathrm{softmax}\left(h_{\texttt{cls}}(\mathbf{z}_{i,j})\right), \qquad \bar{q}_i = \frac{1}{N}\sum_{j=1}^{N} q_{i,j}, \tag{19}$$

where $q_{i,j}$ is the class probability predicted from the j-th semantic patch latent of the i-th image, and $\bar{q}_i$ denotes the image-level prediction.

c) SIGReg-based Independence Loss: To efficiently implement the independence regularizer $L_{\texttt{ind}}$ for NVIB-RAE, we adopt the *Sketched Isotropic Gaussian Regularization* (SIGReg) [84]. SIGReg is a recently proposed form of representation regularizer that encourages a set of embeddings to match an isotropic Gaussian distribution. Instead of directly matching a high-dimensional distribution, SIGReg projects the embeddings onto multiple random directions and applies a one-dimensional Gaussianity test to each projected distribution, for which we use the Epps-Pulley test. This provides a scalable way to encourage isotropic and well-conditioned latent geometry.

Specifically, let $\mathbf{z}_{n,b,j} \in \mathbb{R}^{d_n}$ denote the nuisance latent of the j-th patch token in the b-th image, and let $\mathcal{A}$ be a set of random unit projection directions. Then, we define the SIGReg-based independence loss as the following:

$$L_{\texttt{ind}} = \frac{1}{N|\mathcal{A}|}\sum_{j=1}^{N}\sum_{a\in\mathcal{A}} T_{\mathcal{N}(0,1)}\left(\left\{a^{\top}\mathbf{z}_{n,b,j}\right\}_{b=1}^{B}\right), \tag{20}$$

where $T_{\mathcal{N}(0,1)}$ denotes the Epps-Pulley normality test against the standard Gaussian distribution. For a set of projected samples $U = \{u_b\}_{b=1}^{B}$, it is given by

$$T_{\mathcal{N}(0,1)}(U) = B\int_{-\infty}^{\infty}\left|\widehat{\phi}_U(t) - \phi_{\mathcal{N}(0,1)}(t)\right|^2 w(t)\,dt, \tag{21}$$

where $\widehat{\phi}_U(t) = \frac{1}{B}\sum_{b=1}^{B}\exp(\mathrm{i}tu_b)$ is the empirical characteristic function of the projected samples, $\phi_{\mathcal{N}(0,1)}(t) = \exp(-t^2/2)$ is the characteristic function of the standard Gaussian, and $w(t)$ is the Epps–Pulley weighting function. In practice, the projection directions in $\mathcal{A}$ are random unit vectors, and the loss is averaged over patch locations and projection directions.

Compared with the GAN-based independence objective used for NVIB-VQ in Eq. (16), the SIGReg-based form of $L_{\texttt{ind}}$ provides a scalable independence regularizer for the pretrained representation setting. It matches the projected empirical distributions of deterministic nuisance tokens to a standard Gaussian, thereby encouraging an isotropic and well-conditioned nuisance latent geometry. Together with the VIB regularization on the semantic latent, this promotes an approximately factorized latent space while allowing $\mathbf{z}_n$ to preserve reconstruction-relevant residual information.

D. Training and Inference

The nuisance classifier q_n and the main model are optimized alternately. During nuisance-classifier updates, the nuisance latents are detached so that gradients are propagated only to q_n. During the main optimization step, q_n is fixed while the remaining trainable modules are optimized using the practical NVIB objective in Eq. (10). The overall NVIB training procedure is provided in the supplementary material.

At inference time, only the semantic branch is used for downstream prediction. The nuisance branch and reconstruction modules are auxiliary components used only during training to capture reconstruction-relevant residual variations and encourage semantic-nuisance factorization.

V. Experiments

We evaluate the proposed two NIB implementations, NVIB-RAE and NVIB-VQ, from complementary perspectives on model robustness. We first examine NVIB-RAE in the pretrained representation setting, where large-scale foundation visual encoders [27], [30] are adapted to downstream tasks. In Section V-A, we evaluate its effectiveness on domain generalization and fine-grained recognition. We then turn to NVIB-VQ in Section V-B, showing that the same NIB principle can also be applied to representation learning from scratch. Lastly, Section V-C presents a component-wise ablation study. The detailed experimental setups, *e.g.*, datasets, hyperparameters, etc., can be found in the supplementary material.

A. Evaluation of NVIB-RAE

We first evaluate whether NVIB-RAE remains effective when adapting pretrained visual representations to downstream tasks, primarily in two complementary aspects: (a) domain generalization under natural distribution shifts, and (b) fine-grained transferability.

1) Domain Generalization under Natural Distribution Shifts: To assess the robustness of learned representations under natural distribution shifts, we train on ImageNet-1K [88] and evaluate on naturally shifted ImageNet variants, including ImageNet-A (IN-A) [89], ImageNet-R (IN-R) [38], ImageNet-V2 (IN-V2) [90], and ImageNet-Sketch (IN-Sketch) [91]. As reported in Table I, NVIB-RAE consistently improves domain generalization on both DINOv2-B/14 and SigLIP2-B/16. In the frozen-feature setting, NVIB-RAE improves the average OOD accuracy of DINOv2-B/14 by 5.7%, with notable gains on ImageNet-A and ImageNet-R, where top-1 accuracy increases by 11.0% and 9.7%, respectively. For SigLIP2-B/16, NVIB-RAE improves the average OOD accuracy by 11.3%, with a particularly large top-1 gain of 54.4% on ImageNet-A. The method improves OOD robustness across both DINOv2 and SigLIP2, suggesting that its benefit does not depend on a particular pretrained encoder, but can instead be attributed to its separation of semantic and nuisance information.

The improvement is more pronounced in the LoRA-adapted [92] setting. For DINOv2-B/14, NVIB-RAE improves the average OOD accuracy by 26.2%, with large top-1 gains of 65.4% on ImageNet-A and 52.4% on ImageNet-R. For SigLIP2-B/16, NVIB-RAE improves the average OOD accuracy by 31.2%, with especially large top-1 gains of 101.2% on ImageNet-A and 56.6% on ImageNet-R. Since LoRA introduces task-specific low-rank updates into the pretrained encoder,

TABLE I
COMPARISON OF ACCURACY (%; ↑) FOR DOMAIN GENERALIZATION USING DINOV2 AND SIGLIP2 WITH LINEAR AND LORA ADAPTATION ON IMAGENET. AVG. DENOTES THE AVERAGE TOP-1 ACCURACY OVER THE OOD BENCHMARKS. BOLD INDICATES THE BEST RESULTS.

Method	IN-1K	IN-A		IN-R		IN-V2	IN-Sketch	Avg.
		Top-1	Top-5	Top-1	Top-5			
DINOv2 [27]	**83.0**	54.4	78.4	64.9	77.6	**75.5**	53.5	62.1
+NVIB-RAE (Ours)	**83.0**	**60.4**	**81.0**	**71.2**	**82.0**	75.0	**55.7**	**65.6**
DINOv2 [27] + LoRA	83.1	40.6	67.3	46.8	64.9	75.3	51.8	53.6
+ NVIB-RAE (Ours)	**84.3**	**67.2**	**86.7**	**71.3**	**82.8**	**76.3**	**56.0**	**67.7**
SigLIP2 [30]	79.9	28.8	59.8	78.1	90.6	70.0	59.9	59.2
+NVIB-RAE (Ours)	**81.0**	**44.4**	**77.7**	**85.4**	**95.3**	**71.6**	**62.2**	**65.9**
SigLIP2 [30] + LoRA	82.2	27.9	54.6	54.5	74.5	73.1	58.0	53.4
+ NVIB-RAE (Ours)	**83.6**	**56.1**	**82.7**	**85.3**	**94.7**	**75.3**	**63.4**	**70.0**

TABLE II
COMPARISON OF ACCURACY (%; ↑) FOR FINE-GRAINED BENCHMARKS USING DINOV2 AND SIGLIP2 WITH LINEAR ADAPTATION TO IMAGENET. AVG. DENOTES THE AVERAGE ACCURACY OVER ALL BENCHMARKS. BOLD INDICATES THE BEST RESULTS.

Method	Food	C10.	C100.	SUN	Cars	Aircr.	VOC	DTD	Pets	Cal101	Flowers	CUB	Avg.↑
DINOv2 [27]	92.5	98.6	91.0	78.4	88.3	72.4	**93.3**	**83.8**	94.2	96.8	**99.9**	79.6	89.1
+NVIB-RAE (Ours)	**93.1**	**98.7**	**91.1**	**81.8**	**92.7**	**77.4**	92.5	81.4	**96.0**	**97.9**	**99.9**	**89.2**	**91.0**
SigLIP2 [30]	90.4	93.0	77.2	81.6	93.0	**71.2**	**93.3**	81.0	92.7	**98.2**	**99.8**	73.0	87.0
+NVIB-RAE (Ours)	**94.0**	**96.3**	**84.1**	**82.4**	**93.4**	**71.2**	93.1	**84.0**	**94.7**	96.8	99.5	**80.4**	**89.2**

it can increase reliance on ImageNet-specific cues that are not stable under distribution shifts. For example, in DINOv2-B/14, LoRA slightly improves ImageNet accuracy but decreases the average OOD accuracy by 15.8%. The larger gains in this setting therefore suggest that NVIB-RAE improves robust adaptation by separating pretrained token representations into semantic and nuisance branches, thereby reducing shortcut reliance and promoting more robust transfer.

2) Fine-Grained Transferability: We further investigate whether the semantic latent retains fine-grained class-discriminative information through recognition tasks requiring subtle visual cues. Specifically, we evaluate ImageNet-trained representations on diverse downstream tasks, including Food-101 [93], CIFAR-10/CIFAR-100 [94], SUN397 [95], Stanford Cars [96], FGVC-Aircraft [97], Pascal VOC [98], DTD [99], Oxford-IIIT Pets [100], Caltech-101 [101], Oxford Flowers-102 [102], and CUB-200-2011 [103]. NVIB-RAE consistently improves fine-grained transferability across diverse downstream recognition benchmarks, as shown in Table II. For DINOv2-B/14, NVIB-RAE improves the average accuracy by 2.1%, with particularly large gains on CUB, Aircraft, Cars, and SUN, where accuracy increases by 12.0%, 7.0%, 5.0%, and 4.5%, respectively. For SigLIP2-B/16, NVIB-RAE improves the average accuracy by 2.5%, with notable gains on CUB and CIFAR-100, where accuracy increases by 10.1% and 8.9%, respectively. These results indicate that the semantic latent preserves fine-grained discriminative information, rather than merely improving robustness through aggressive compression. Thus, NVIB-RAE maintains transferable semantic details while separating nuisance variation from the representation.

TABLE III
COMPARISON OF ERROR RATES (%; ↓) ON CIFAR-10 AND ITS VARIANTS. BOLD AND UNDERLINE INDICATE THE BEST AND RUNNER-UP, RESPECTIVELY. †PIXMIX [54] UTILIZES AN EXTERNAL DATASET.

Method	C10	C10-C	C10.1	C10.2	CINIC
Cross-entropy	6.08	16.0	13.4	18.3	23.7
VIB [72]	5.98	15.2	13.6	16.8	23.6
NLIB [104]	6.81	17.0	14.6	17.5	24.3
sq-NLIB [105]	6.02	15.5	13.0	17.1	23.7
DisenIB [73]	5.76	15.2	13.2	17.2	23.7
AugMix [37]	6.52	15.1	14.2	17.2	24.2
PixMix† [54]	5.43	<u>10.3</u>	13.1	16.6	23.2
NVIB-VQ (Ours)	<u>4.97</u>	12.3	<u>11.6</u>	<u>15.5</u>	<u>22.2</u>
+ AugMix [37]	5.35	12.0	12.5	15.8	22.6
+ PixMix† [54]	**4.67**	**8.08**	**10.4**	**14.8**	**22.1**

B. Evaluation of NVIB-VQ

In this section, we assess NVIB-VQ to examine whether NIB remains effective when representations are learned from scratch, rather than adapted from pretrained encoders. Here, we test two distinct aspects: (a) corruption and natural robustness, and (b) novelty detection. Both these aspects have been considered challenging to improve for models trained from scratch, especially in the absence of task-specific priors.

1) Robustness against Natural Corruptions: We evaluate whether NVIB-VQ improves robustness to natural corruptions and distribution shifts when representations are learned from scratch. Specifically, we train models on CIFAR-10 [94] and ImageNet [88] and compare performances on their distribution-shifted variants.

Table III reports results on CIFAR-based benchmarks. Compared with the cross-entropy baseline, NVIB-VQ reduces

TABLE IV
COMPARISON OF ERROR RATES (%; ↓) OR MEAN CORRUPTION ERROR (MCE, %; ↓) ON IMAGENET AND ITS VARIANTS.

	ViT-S/16		ViT-B/16	
Dataset	Baseline	**NVIB-VQ**	Baseline	**NVIB-VQ**
IN-1K [88]	**25.1**	**25.1**	**21.8**	21.9
IN-C (mCE) [18]	65.9	**65.2** (−0.7)	58.6	**57.5** (−1.1)
IN-R [38]	70.3	**67.1** (−3.2)	66.3	**64.4** (−1.9)
IN-Sketch [91]	80.3	**77.7** (−2.6)	76.5	**74.4** (−2.1)

TABLE V
EVALUATION ON BACKGROUNDS CHALLENGE [106], COMPARED TO THE CROSS-ENTROPY BASELINE. ALL THE MODELS ARE TRAINED ON IMAGENET, AND WARPED TO PERFORM CLASSIFICATION ON IMAGENET-9.

BG-Challenge	ViT-S/16		ViT-B/16	
Dataset	Baseline	**NVIB-VQ**	Baseline	**NVIB-VQ**
ORIGINAL (IN-9; ↑)	95.3	**95.5**	96.0	**96.1**
ONLY-BG-T (↓)	20.3	**17.8** (−2.5)	24.2	**21.1** (−3.1)
MIXED-SAME (↑)	86.3	**88.3** (+2.0)	87.4	**88.9** (+1.5)
MIXED-RAND (↑)	77.8	**80.5** (+2.7)	80.1	**81.8** (+1.7)
BG-gap (↓)	8.5	**7.8** (−0.7)	7.3	**7.1** (−0.2)

the CIFAR-10 error by 18.3% and the CIFAR-10-C [18] error by 23.1%. It also improves robustness on natural shift benchmarks, reducing the error by 13.4% on CIFAR-10.1, 15.3% on CIFAR-10.2, and 6.3% on CINIC-10. Notably, NVIB-VQ achieves lower errors than PixMix [54] on CIFAR-10, CIFAR-10.1, CIFAR-10.2, and CINIC-10, even though PixMix uses an external dataset of pattern- and fractal-like images. When combined with PixMix, NVIB-VQ further reduces the CIFAR-10-C error by 21.6%, showing that nuisance modeling is complementary to augmentation-based robustness methods.

Table IV further shows that the robustness benefit of NVIB-VQ extends to ImageNet. For ViT-S/16, NVIB-VQ preserves the ImageNet validation error while reducing the error on ImageNet-C [18], ImageNet-R [38], and ImageNet-Sketch [91] by 1.1%, 4.6%, and 3.2%, respectively. For ViT-B/16, NVIB-VQ also improves all shifted benchmarks, reducing the error by 1.9% on ImageNet-C, 2.9% on ImageNet-R, and 2.7% on ImageNet-Sketch, with nearly unchanged ImageNet error rate. These results show that NVIB-VQ remains effective beyond small-scale CIFAR benchmarks. Overall, these findings suggest that explicitly modeling nuisance information helps the representation become less sensitive to corruption and distribution shifts. Unlike methods that rely mainly on predefined augmentations or external auxiliary data, NVIB-VQ improves robustness by separating semantic information from residual nuisance variation.

Lastly, Table V evaluates ImageNet models on the Background Challenge [106], a benchmark to assess robustness to background shifts. It builds multiple variants of ImageNet-9, which groups 370 ImageNet subclasses, by systematically changing the background composition. Across this benchmark, NVIB-VQ consistently outperforms the cross-entropy baseline, indicating that NVIB-VQ encourages the model to learn features that are less biased toward background cues.

2) Novelty Detection: Given that NIB learns generative representations through nuisance modeling, we further investigate whether these representations can also benefit novelty detection, *i.e.*, identifying unseen inputs based on internal representations of the model. Unlike the standard maximum confidence score, *e.g.*, $\max_y p(y|x)$, NVIB-VQ can additionally use the nuisance likelihood $\log \mathcal{N}(z_n; 0, I) = -\frac{1}{2}\|\mathbf{z}_n\|^2$, which measures whether the nuisance latent follows the learned in-distribution nuisance structure.

In addition, to detect harder novelties, we replace $\max_y p(y|x)$ with the log-likelihood of $p(y|x) \in \Delta^{|\mathcal{Y}|-1}$ under a symmetric Dirichlet distribution. With a concentration parameter $\alpha > 0$, we have:

$$\log \mathrm{Dir}_\alpha(p(y|x)) = (\alpha - 1) \sum_{k=1}^{|\mathcal{Y}|} \log p_k(y|x) + \text{const.} \quad (22)$$

As $\alpha \to 0$, this distribution approaches a symmetric one-hot distribution, which is well aligned with the behavior encouraged by standard classification objectives. We simply set $\alpha = 0.05$ throughout the experiments.

Table VI reports the performance of the combined score, $\log \mathrm{Dir}_{0.05}(y) + \log \mathcal{N}(z_n; 0, I)$ on the OBJECTS benchmark [48], which features a challenging full-spectrum OOD detection setting with both covariate and semantic shifts as in-distribution. Compared with SEM scoring, the nuisance-aware score improves AUROC by 22.1% on MNIST, 6.9% on FashionMNIST, 11.6% on Texture, and 4.8% on CIFAR-100-C. These improvements suggest that semantic confidence alone is insufficient for reliable OOD detection, while the nuisance latent provides complementary information by capturing residual variations overlooked by the semantic representation.

C. Ablation Study

We further conduct an ablation study to examine the individual contributions of the components in NVIB. Unless otherwise noted, the results reported in this section are based on the NVIB-VQ trained on CIFAR-10, following the training details specified in the supplementary material.

1) Reconstruction Loss: The reconstruction loss $L_{\texttt{recon}}$ is an essential component for making NVIB-VQ work as a nuisance model, since it prevents residual information from being discarded and forces $(\mathbf{z}, \mathbf{z}_n)$ to preserve the input. When this component is ablated, by removing $L_{\texttt{recon}}$ from the NVIB-VQ training objective, we observe that the overall training becomes unstable and results in degradation in both accuracy and corruption robustness, as shown in Table VII. This confirms that reconstruction is necessary for learning a robust semantic representation in NVIB.

2) Nuisance Loss: From the ablation of $L_{\texttt{nuis}}$ in Table VII, we observe that removing this loss degrades both clean accuracy and corruption robustness. This indicates that explicitly enforcing the nuisance property of $\mathbf{z}_n$ with respect to $\mathbf{y}$ is important for learning a robust semantic representation $\mathbf{z}$. By making $\mathbf{z}_n$ preserve residual information needed to infer $\mathbf{x}$ while suppressing its label-discriminative information, $L_{\texttt{nuis}}$ encourages $\mathbf{z}$ to capture class-related information more reliably.

TABLE VI
COMPARISON OF OOD DETECTION PERFORMANCE ON THE OBJECTS BENCHMARK [48], WHICH CONSIDERS CIFAR-10-C AND IMAGENET-10 AS IN-DISTRIBUTION AS WELL AS THE TRAINED CIFAR-10. BOLD AND UNDERLINE DENOTE THE BEST AND RUNNER-UP, RESPECTIVELY.

FS-OOD: OBJECTS		AUROC (%; ↑) / AUPR (%; ↑) / FPR@TPR95 (%; ↓)			
Method	Score	MNIST	FashionMNIST	Texture	CIFAR-100-C
Cross-entropy	$\max_y p(y\|x)$ [43]	66.98 / 52.66 / 93.54	73.78 / 90.15 / 88.08	74.18 / 93.34 / 85.64	74.12 / 89.74 / 87.26
	ODIN [107]	70.31 / 49.58 / 82.04	80.98 / 91.53 / **68.73**	70.14 / 89.97 / <u>72.91</u>	67.51 / 83.97 / 84.26
	Energy-based [45]	54.55 / 34.14 / 92.23	76.50 / 89.80 / <u>72.40</u>	68.63 / 89.51 / 75.57	68.37 / 85.54 / 83.64
	Mahalanobis [44]	77.04 / 65.31 / 84.59	80.33 / 92.28 / 77.17	72.02 / 88.46 / 72.98	68.13 / 82.97 / 85.53
	SEM [48]	75.69 / 76.61 / 99.70	79.40 / 93.14 / 93.72	<u>79.69</u> / <u>95.48</u> / 82.15	78.89 / 92.07 / 83.92
	$\log \mathbf{Dir}_{0.05}(\boldsymbol{y})$	76.75 / 66.26 / 83.51	82.88 / 93.97 / 77.19	70.69 / 92.68 / 91.35	78.80 / 92.21 / 82.50
VIB [72]	$\max_y p(y\|x)$ [43]	80.23 / 73.50 / 80.69	76.35 / 91.22 / 84.75	74.67 / 94.09 / 87.22	76.12 / 91.03 / 84.99
	$\log \mathbf{Dir}_{0.05}(\boldsymbol{y})$	86.13 / 79.45 / 64.92	81.11 / 93.12 / 77.82	73.84 / 93.50 / 88.00	78.54 / 91.85 / <u>81.47</u>
NVIB-VQ (ours)	$\max_y p(y\|x)$ [43]	79.67 / 71.50 / 80.22	77.33 / 91.63 / 84.31	74.95 / 93.97 / 86.01	74.31 / 89.89 / 86.26
	$\log \mathbf{Dir}_{0.05}(\boldsymbol{y})$	<u>90.53</u> / <u>85.68</u> / <u>52.08</u>	<u>84.56</u> / <u>94.61</u> / 74.24	75.04 / 93.83 / 86.01	<u>79.39</u> / <u>92.33</u> / 81.51
	$+\log \mathcal{N}(\boldsymbol{z}_n; \mathbf{0}, \boldsymbol{I})$	**92.43** / **89.38** / **48.10**	**84.85** / **94.84** / 74.67	**88.91** / **97.49** / **48.44**	**82.66** / **93.62** / **74.14**

TABLE VII
COMPARISON OF THE TEST ERROR RATE (ERR.; %, ↓), CORRUPTION ERROR (C-ERR.; %, ↓) ON CIFAR-10 [94] ACROSS NVIB-VQ ABLATIONS. WE TRAIN VIT-S/4 FOR THIS EXPERIMENT.

$L_{\texttt{recon}}$	$L_{\texttt{nuis}}$	$L_{\texttt{ind}}$	KL	Err.	C-Err.
✓	✓	✓	✓	10.9	22.0
✗	✓	✓	✓	52.6	60.3
✓	✗	✓	✓	12.4	24.9
✓	✓	✗	✓	11.4	22.6
✓	✓	✓	✗	11.3	22.8

Fig. 3. Reconstructions under random nuisance $\mathbf{z}_n$ (upper rows) and random semantic $\mathbf{z}$ (lower row). The leftmost per row shows the original reconstruction.

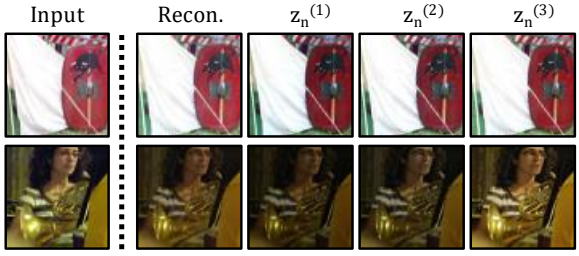


Fig. 4. Qualitative analysis of learned nuisances. Each image is reconstructed from latent where (a) the semantic part is fixed from the original and (b) the nuisance part is replaced with those from other samples.

3) Independence Loss: The independence loss $L_{\texttt{ind}}$ performs adversarial prior matching so that the joint latent distribution of the semantic and nuisance variables follows a factorized Gaussian prior, *i.e.*, $p(\mathbf{z}, \mathbf{z}_n) \sim \mathcal{N}(0, \mathbf{I})$. This encourages $\mathbf{z} \perp \mathbf{z}_n$ and makes the latent space more tractable to sample. Although this component is primarily for generative modeling (and its applications, *e.g.*, to novelty detection), removing $L_{\texttt{ind}}$ also degrades the final clean accuracy and corruption robustness, as shown in Table VII, suggesting the effect of $L_{\texttt{ind}}$ based on Lemma 1 for the semantic-nuisance separation.

4) KL Regularization: The KL regularization term in VIB controls the capacity of the semantic latent $\mathbf{z}$ by encouraging its posterior distribution to stay close to the prior. In NVIB training, this regularization is important for balancing the information allocation between the semantic $\mathbf{z}$ and the nuisance $\mathbf{z}_n$. Without the KL regularization, $\mathbf{z}$ can encode excessive residual information, which weakens the intended semantic-nuisance separation. As shown in Table VII, removing this term increases both the standard test error and the corruption error, indicating that the KL regularization is necessary for learning a robust semantic representation.

5) Qualitative Analysis of Learned Nuisances: We further provide a qualitative nuisance randomization analysis to inspect whether the learned semantic and nuisance representations behave as intended. Figure 3 examines how different values of β affect the learned representations, by comparing the reconstructed samples for a fixed input while randomizing the nuisance $\mathbf{z}_n$. We observe a clear trend from this comparison demonstrating the effect of β: having a larger β makes the model push more semantic information into $\mathbf{z}_n$ regarding it as the nuisance. Without the information bottleneck, *i.e.*, in the case when $\beta = 0.0$, we qualitatively observe that the model rather encodes most information in $\mathbf{z}$, due to the minimax loss applied to the nuisance $\mathbf{z}_n$. As shown in Figure 4, NVIB-RAE first reconstructs each input using its own semantic latent $\mathbf{z}$ and nuisance latent $\mathbf{z}_n$. We then fix $\mathbf{z}$ and replace $\mathbf{z}_n$ with nuisance latents from other samples. The resulting reconstructions preserve the main semantic content of the input while altering residual visual factors. This suggests that $\mathbf{z}$ captures prediction-relevant semantic information, whereas $\mathbf{z}_n$

accounts for nuisance variation.

VI. Conclusion

In this work, we demonstrate that learning an effective nuisance model provides a practical approach to inducing robust representations. Building upon the principle of NIB, we develop two complementary implementations applicable to both modern representation learning paradigms: NVIB-VQ for end-to-end learning from scratch and NVIB-RAE for adapting pretrained visual representations. Across diverse downstream settings, the proposed frameworks consistently improve robustness, generalization, transferability, and novelty detection, notably without relying on task-specific robustness priors. We believe this work represents a meaningful step toward understanding how nuisance modeling contributes to out-of-distribution generalization. More broadly, we hope this work motivates a new perspective on representation learning: robustness should emerge not from discarding nuisance information, but from understanding and modeling it explicitly.

References

[1] A. Brock, S. De, S. L. Smith, and K. Simonyan, "High-performance large-scale image recognition without normalization," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1059–1071.

[2] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[3] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr *et al.*, "SAM 3: Segment anything with concepts," 2025.

[4] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[6] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

[7] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[8] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first International Conference on Machine Learning*, 2024.

[9] Y. Yu, W. Xiong, W. Nie, Y. Sheng, S. Liu, and J. Luo, "Pixeldit: Pixel diffusion transformers for image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[10] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *European Conference on Computer Vision*, 2020.

[11] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5470–5479.

[12] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions on Graphics*, vol. 41, no. 4, pp. 102:1–102:15, 2022.

[13] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 139:1–139:14, 2023.

[14] Z. Fan, W. Cong, K. Wen, K. Wang, J. Zhang, X. Ding, D. Xu, B. Ivanovic, M. Pavone, G. Pavlakos *et al.*, "Instantsplat: Sparse-view gaussian splatting in seconds," *arXiv preprint arXiv:2403.20309*, 2024.

[15] Wan Team, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.

[16] S. Yuan, Y. Yin, Z. Li, X. Huang, X. Yang, and L. Yuan, "Helios: Real real-time long video generation model," *arXiv preprint arXiv:2603.04379*, 2026.

[17] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2014.

[18] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *International Conference on Learning Representations*, 2019.

[19] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 853–17 862.

[20] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 164–10 183, 2024.

[21] L. Lan, P. Cai, L. Jiang, X. Liu, Y. Li, and Y. Zhang, "Brau-net++: U-shaped hybrid cnn-transformer network for medical image segmentation," *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2026.

[22] O. F. Atli, B. Kabas, F. Arslan, A. C. Demirtas, M. Yurt, O. Dalmaz, and T. Cukur, "I2I-Mamba: Multi-modal medical image synthesis via selective state space modeling," *IEEE Transactions on Biomedical Engineering*, 2026.

[23] E. Asgari, N. Montaña-Brown, M. Dubois, S. Khalil, J. Balloch, J. A. Yeung, and D. Pimenta, "A framework to assess clinical safety and hallucination rates of llms for medical text summarisation," *NPJ digital medicine*, vol. 8, no. 1, p. 274, 2025.

[24] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[26] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9650–9660.

[27] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "DINOv2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[28] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski, "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025.

[29] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 975–11 986.

[30] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa *et al.*, "SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," *arXiv preprint arXiv:2502.14786*, 2025.

[31] Y. Gao, C. Chen, and J. Gu, "One layer is enough: Adapting pretrained visual encoders for image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 4688–4697.

[32] M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. G. Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong *et al.*, "Robust fine-tuning of zero-shot models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7959–7971.

[33] S. Goyal, A. Kumar, S. Garg, Z. Kolter, and A. Raghunathan, "Finetune like you pretrain: Improved finetuning of zero-shot vision models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 338–19 347.

[34] Anonymous Authors, "Enhancing multiple reliability measures via nuisance-extended information bottleneck," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[35] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[36] B. Zheng, N. Ma, S. Tong, and S. Xie, "Diffusion transformers with representation autoencoders," *arXiv preprint arXiv:2510.11690*, 2025.

[37] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: A simple method to improve robustness and uncertainty under data shift," in *International Conference on Learning Representations*, 2020.

[38] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2021, pp. 8340–8349.

[39] S. S. Latif, S. Shiper, K. Kiran, M. F. Ishmam, M. A. Hossain, A. R. M. Kamal, and M. H. Ashmafee, "Enhancing vision language corruption robustness using cross-distribution & prompted denoisers," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026, pp. 5754–5765.

[40] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[41] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," 2019.

[42] E.-C. Chen, P.-Y. Chen, I. Chung, C.-R. Lee *et al.*, "Data-driven lipschitz continuity: A cost-effective approach to improve adversarial robustness," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026, pp. 698–707.

[43] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *International Conference on Learning Representations*, 2017.

[44] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018.

[45] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 464–21 475, 2020.

[46] Z. Xu, Q. Xu, Z. Wang, C. Hua, S. Li, Z. Yang, and Q. Huang, "Mind the way you select negative texts: Pursuing the distance consistency in OOD detection with VLMs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 34 650–34 661.

[47] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," in *International Conference on Learning Representations*, 2019.

[48] J. Yang, K. Zhou, and Z. Liu, "Full-spectrum out-of-distribution detection," *International Journal of Computer Vision*, vol. 131, no. 10, pp. 2607–2622, 2023.

[49] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. Earnshaw, I. Haque, S. M. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang, "WILDS: A benchmark of in-the-wild distribution shifts," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 5637–5664.

[50] J. Zhang, J. Yang, P. Wang, H. Wang, Y. Lin, H. Zhang, Y. Sun, X. Du, K. Zhou, W. Zhang, Y. Li, Z. Liu, Y. Chen, and H. Li, "OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection," *Journal of Data-centric Machine Learning Research*, 2024.

[51] M. Koddenbrock, R. Hoffmann, D. Brodmann, and E. Rodner, "On the domain robustness of contrastive vision-language models," in *German Conference on Artificial Intelligence (Künstliche Intelligenz)*. Springer, 2025, pp. 62–76.

[52] Z. Xu, X. Xiang, and Y. Liang, "Overcoming shortcut problem in vlm for robust out-of-distribution detection," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 15 402–15 412.

[53] S. Wang, R. Veldhuis, and N. Strisciuglio, "Do imagenet-trained models learn shortcuts? the impact of frequency shortcuts on generalization," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 25 198–25 207.

[54] D. Hendrycks, A. Zou, M. Mazeika, L. Tang, B. Li, D. Song, and J. Steinhardt, "PixMix: Dreamlike pictures comprehensively improve safety measures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 783–16 792.

[55] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, "Vision transformers need registers," in *International Conference on Learning Representations*, 2024.

[56] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 38–55.

[57] S. Yu, S. Kwak, H. Jang, J. Jeong, J. Huang, J. Shin, and S. Xie, "Representation alignment for generation: Training diffusion transformers is easier than you think," in *International Conference on Learning Representations*, 2025.

[58] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," in *International Conference on Learning Representations*, 2022.

[59] Z. Li, C. Chen, T. Xu, Z. Qin, J. Xiao, Z.-Q. Luo, and R. Sun, "Preserving diversity in supervised fine-tuning of large language models," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 66 127–66 154.

[60] Y. Lee, A. S. Chen, F. Tajwar, A. Kumar, H. Yao, P. Liang, and C. Finn, "Surgical fine-tuning improves adaptation to distribution shifts," in *International Conference on Learning Representations*, 2023.

[61] G. Nam, S. Lee, J.-Y. Lee, and H. J. Kim, "Lipsum-ft: Robust fine-tuning of zero-shot models using random text guidance," in *International Conference on Learning Representations*, 2024.

[62] S. Kim, H. Kim, M. Cho, and S. Kwak, "Efficient and versatile robust fine-tuning of zero-shot models," in *European Conference on Computer Vision*, 2024.

[63] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky, "Your classifier is secretly an energy based model and you should treat it like one," in *International Conference on Learning Representations*, 2020.

[64] A. Li, A. Kumar, and D. Pathak, "Generative classifiers avoid shortcut solutions," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 15 932–15 952.

[65] X. Yin, C. Zhang, J. Steele, N. N. Shavit, and T. T. Wang, "Scalable energy-based models via adversarial training: Unifying discrimination and generation," in *The Fourteenth International Conference on Learning Representations*, 2026.

[66] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan, "Do deep generative models know what they don't know?" in *International Conference on Learning Representations*, 2019.

[67] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, and B. Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[68] J. Serra, D. Alvarez, V. Gomez, O. Slizovskaia, J. F. Nunez, and J. Luque, "Input complexity and out-of-distribution detection with likelihood-based generative models," in *International Conference on Learning Representations*, 2020.

[69] H. Kamkari, B. L. Ross, J. C. Cresswell, A. L. Caterini, R. G. Krishnan, and G. Loaiza-Ganem, "A geometric explanation of the likelihood OOD detection paradox," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 235. PMLR, 2024.

[70] X. Yang and S. Ji, "JEM++: Improved techniques for training JEM," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2021, pp. 6494–6503.

[71] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," *arXiv preprint arXiv:1503.02406*, 2015.

[72] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[73] Z. Pan, L. Niu, J. Zhang, and L. Zhang, "Disentangled information bottleneck," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, 2021, pp. 9285–9293.

[74] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[75] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[76] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, "Learning not to learn: Training deep neural networks with biased data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9012–9020.

[77] K. Ahuja, E. Caballero, D. Zhang, J.-C. Gagnon-Audet, Y. Bengio, I. Mitliagkas, and I. Rish, "Learning invariant representations using inverse contrastive loss," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[78] B. Li, Y. Shen, Y. Wang, W. Zhu, C. J. Reed, J. Zhang, D. Li, K. Keutzer, and H. Zhao, "Invariant information bottleneck for domain generalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[79] K. Ahuja, E. Caballero, D. Zhang, J.-C. Gagnon-Audet, Y. Bengio, I. Mitliagkas, and I. Rish, "Invariance principle meets information bottleneck for out-of-distribution generalization," *arXiv preprint arXiv:2106.06607*, 2021.

[80] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[81] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[82] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2014.

[83] M. Chalk, O. Marre, and G. Tkacik, "Relevant sparse codes with variational information bottleneck," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016.

[84] R. Balestriero and Y. LeCun, "LeJEPA: Provable and scalable self-supervised learning without the heuristics," *arXiv preprint arXiv:2511.08544*, 2025.

[85] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[86] A. van den Oord, O. Vinyals, and k. kavukcuoglu, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[87] J. Yu, X. Li, J. Y. Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge, and Y. Wu, "Vector-quantized image modeling with improved VQGAN," in *International Conference on Learning Representations*, 2022.

[88] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[89] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 262–15 271.

[90] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do ImageNet classifiers generalize to ImageNet?" in *International Conference on Machine Learning*. PMLR, 2019, pp. 5389–5400.

[91] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning robust global representations by penalizing local predictive power," in *Advances in Neural Information Processing Systems*, 2019, pp. 10 506–10 518.

[92] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[93] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – mining discriminative components with random forests," in *European Conference on Computer Vision*, 2014, pp. 446–461.

[94] A. Krizhevsky, "Learning multiple layers of features from tiny images," Department of Computer Science, University of Toronto, Tech. Rep., 2009.

[95] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3485–3492.

[96] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 554–561.

[97] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," University of Oxford, Tech. Rep., 2013.

[98] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[99] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3606–3613.

[100] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3498–3505.

[101] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[102] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Indian Conference on Computer Vision, Graphics & Image Processing*, 2008, pp. 722–729.

[103] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[104] A. Kolchinsky, B. D. Tracey, and D. H. Wolpert, "Nonlinear information bottleneck," *Entropy*, vol. 21, no. 12, p. 1181, 2019.

[105] B. Rodríguez Gálvez, R. Thobaben, and M. Skoglund, "The convex information bottleneck Lagrangian," *Entropy*, vol. 22, no. 1, p. 98, 2020.

[106] K. Y. Xiao, L. Engstrom, A. Ilyas, and A. Madry, "Noise or signal: The role of image backgrounds in object recognition," in *International Conference on Learning Representations*, 2021.

[107] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *International Conference on Learning Representations*, 2018.