ObjectCompose: Evaluating Resilience of Vision-Based Models on Object-to-Background Compositional Changes

Hashmat Shadab Malik*[1], Muhammad Huzaifa*[1], Muzammal Naseer[2], Salman Khan[1,3], and Fahad Shahbaz Khan[1,4]

[1] Mohamed bin Zayed University of AI
[2] Center of Secure Cyber-Physical Security Systems
[3] Australian National University
[4] Linköping University
hashmat.malik@mbzuai.ac.ae, muhammad.huzaifa@mbzuai.ac.ae,
muhammadmuzammal.naseer@ku.ac.ae, salman.khan@mbzuai.ac.ae,
fahad.khan@mbzuai.ac.ae

Abstract. Given the large-scale multi-modal training of recent vision-based models and their generalization capabilities, understanding the extent of their robustness is critical for their real-world deployment. In this work, our goal is to evaluate the resilience of current vision-based models against diverse object-to-background context variations. The majority of robustness evaluation methods have introduced synthetic datasets to induce changes to object characteristics (viewpoints, scale, color) or utilized image transformation techniques (adversarial changes, common corruptions) on real images to simulate shifts in distributions. Recent works have explored leveraging large language models and diffusion models to generate changes in the background. However, these methods either lack in offering control over the changes to be made or distort the object semantics, making them unsuitable for the task. Our method, on the other hand, can induce diverse object-to-background changes while preserving the original semantics and appearance of the object. To achieve this goal, we harness the generative capabilities of text-to-image, image-to-text, and image-to-segment models to automatically generate a broad spectrum of object-to-background changes. We induce both natural and adversarial background changes by either modifying the textual prompts or optimizing the latents and textual embedding of text-to-image models. This allows us to quantify the role of background context in understanding the robustness and generalization of deep neural networks. We produce various versions of standard vision datasets (ImageNet, COCO), incorporating either diverse and realistic backgrounds into the images or introducing color, texture, and adversarial changes in the background. We conduct thorough experimentation and provide an in-depth analysis of the robustness of vision-based models against object-to-background context variations across different tasks. Our code and evaluation benchmark will be available at https://github.com/Muhammad-Huzaifaa/ObjectCompose.

* Equal contribution.

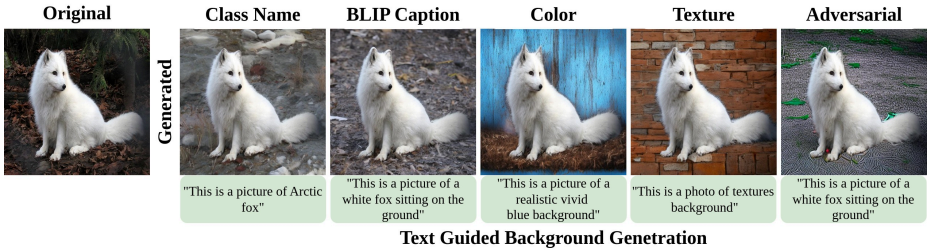

Fig. 1: Image-to-background variations generated by our method, with each column representing a specific background based on the prompt below.

Keywords: Robustness · Adversarial · Foundation Models

1 Introduction

Deep learning-based vision models have achieved significant improvement in diverse vision tasks. However, the performance on static held-out datasets does not capture the diversity of different object background compositions present in the real world. Previous works have shown that vision models are vulnerable to a variety of image alterations, including common corruptions (e.g., snow, fog, blur) [20, 38], domain shifts (e.g., paintings, sketches, cartoons) [18, 19], and changes in viewpoint (e.g., pose, shape, orientation) [3, 5, 26]. Additionally, carefully designed perturbations can be added to images to create adversarial examples that are imperceptible to humans but can fool the decision-making of vision models [16, 51].

Several approaches have been proposed to improve the out-of-distribution robustness of vision models. To achieve adversarial robustness, models are typically trained on adversarial examples [37], and various augmentation policies were proposed to improve non-adversarial robustness of models [8, 19, 53, 55]. More recently, the computer vision field has seen the emergence of large-scale pre-training of both vision [29, 40] and vision-language models [31, 42, 50]. Trained on large-scale datasets and multiple modalities, these models demonstrate promising performance on non-adversarial distribution shifts. Consequently, several works [28, 57] have adapted these models for downstream tasks by utilizing learnable prompts to preserve the rich feature space learned during pre-training.

To evaluate the vision-based models on different distribution shifts, numerous benchmarking datasets, comprising either synthetic or altered real images have been proposed. While synthetic datasets [5, 15, 27] offer more control over desired changes (background, shape, size, viewpoint), most of them capture only simple shape objects within a controlled environment. On the other hand, several studies [20, 38] employ coarse-grained image manipulations on the available ImageNet dataset [10]. However, the coarse-grained transformations used do not encompass the diverse changes that can be induced in real images.

The main motivation of our work is to understand how object-to-background compositional changes in the scene impact uni/multi-modal model performance. Recent works [33, 41] have focused on leveraging existing foundational models to forge new ways to evaluate the resilience of uni/multi-modal vision models. In [41], large language models and text-to-image diffusion models are used for generating diverse semantic changes in real images. However, their method employs the prompt-to-prompt method [22] for image editing, allowing limited word changes in the textual prompt to preserve object semantics. We also observe that it suffers from object distortion due to absence of strong guidance between the object and background during image editing. In [33], diffusion models are used for background editing in real images, and ImageNet-E(diting) dataset is introduced for benchmarking. However, their use of a frequency-based loss for guiding the generation process of diffusion models limit the control to attribute changes in the background. This imposes limitations on the type of background-compositions that can be achieved.

In this work, we develop a framework to investigate the resilience of vision models to diverse object-to-background changes. Leveraging the complementary strengths of image-to-text [31], image-to-segment [29], and text-to-image [24, 43] models, our approach better handles complex background variations. We preserve object semantics (Figure 1) by conditioning the text-to-image diffusion model on object boundaries and textual descriptions from image-to-segment and image-to-text models. We guide the diffusion model by adding the desired textual description or optimizing its latent visual representation and textual embedding for generating diverse natural and adversarial background changes. Additionally, we produce datasets with varied backgrounds from subsets of ImageNet [10] and COCO [34], facilitating the evaluation of uni-modal and multi-modal models. Our contributions are as follows:

- We propose ObjectCompose, an automated approach to introduce diverse background changes to real images, allowing us to evaluate resilience of modern vision-based models against object-to-background context.
- Our proposed background changes yield an average performance drop of 13.64% on classifier models compared to the baseline method, and a substantial drop of 68.71% when exposed to adversarial changes (see Table 1).
- Object detection and segmentation models, which incorporate object-to-background context, display reasonably better robustness to background changes than classification models (see Table 5 and Figure 8).
- Models trained on large-scale datasets with scalable and stable training show better robustness against background changes (see Figure 7 and Table 2).

2 Related Work

Common Corruptions. In [58], different datasets are curated by separating foreground and background elements using ImageNet-1k bounding boxes. They found that models could achieve high object classification performance even when the actual object was absent. Similarly, [44] demonstrate that subtle

changes in object positioning could significantly impact the detector's predictions, highlighting the sensitivity of these models to spatial configurations. A related approach by [48] focuses on co-occurring objects within an image and investigates if removing one object affected the response of the target model toward another. [52] analyze the models' reliance on background signals for decision-making by training on various synthetic datasets. [20] benchmark the robustness of classifiers against common corruptions and perturbations like fog, blur, and contrast variations. In subsequent work, [21] introduce ImageNet-A dataset, filtering natural adversarial examples from a subset of ImageNet to limit spurious background cues. Also, [19] introduce the ImageNet-R dataset, which comprises various renditions of object classes under diverse visual representations such as paintings, cartoons, embroidery, sculptures, and origami. Similarly, [38] introduce the RIVAL10 dataset to study Gaussian noise corruptions in the foreground, background, and object attributes.

Viewpoint Changes. [1,5,15] introduce a large-scale 3D shape datasets to study object scale and viewpoints variations. In a similar vein, [27] introduce a synthetic dataset of rendered objects to aid in diagnostic evaluations of visual question-answering models. Later works have made strides in addressing the realism gap. [2] utilize crowd-sourcing to control rotation, viewpoints, and backgrounds of household objects, while [26] provide more fine-grained annotations for variations on the ImageNet validation set. In a recent development, [3] released PUG dataset rendered using Unreal Engine under diverse conditions, including varying sizes, backgrounds, camera orientations, and light intensities. While these methods offer control over changing several attributes in images, they lack in realism and are not suitable for our primary goal of studying object-to-background context in real images. In contrast, our proposed framework can generate a wide range of object-to-background compositional changes that can influence the models performance.

Adversarial and Counterfactual Manipulations. Researchers have uncovered that subtle, carefully designed alterations to an image, imperceptible to human observers, have the ability to deceive deep learning models [16,30,51]. These perturbations, constructed using gradient-based methods, serve as a worst-case analysis in probing the model's robustness within specified distance norm metrics (l_2 or l_∞). Another strategy entails applying unbounded perturbations to specific image patches, thereby conserving object semantics while inducing model confusion [12,47]. Several studies leverage generative models to create adversarial alterations in images. [6,7] utilise diffusion model and GANs to introduce global adversarial perturbations in the image with strong constraint to semantic changes in order to preserve the original layout of the scene. More recent works [33,41] are more closely related to our goal of evaluating vision-based models on object-to-background compositional changes in the scene. In [41], LANCE framework is proposed which utilises fine-tuned large language models to get the modified textual prompt for editing of attributes in real images. However, this framework is not ideal for studying object-to-background compositional changes since the prompt-to-prompt [22] based image editing method often

leads to global changes in the scene, often altering the object semantics. This necessitates hyper-parameter tuning of parameters used for prompt-to-prompt editing, leading to generation of multiple edited image versions and selecting the one most faithful to the original image. In [33], using already available masks of ImageNet dataset [13], diffusion model is utilised to alter the background of images by varying its texture. A complexity loss based on gray-level co-occurence matrix [17] of the image is used during the denoising process to vary the complexity of the background. A concurrent work [54] evaluates the resilience of models on synthetic images where both the object and background are generated using a diffusion model. In contrast to previous works, our method induces natural/adversarial background variations in real images through textual guidance and optimization of latent space of the diffusion model, all while maintaining the integrity of the object semantics. Our method can be applied to standard vision datasets to generate diverse background variations, providing a robust benchmark for evaluating vision-based models.

3 Method

We introduce ObjectCompose, a method for generating diverse language-guided object-to-background compositional changes to evaluate the resilience of vision models. ObjectCompose leverages the complementary strengths of image-to-segment and image-to-text models to guide object-preserving diffusion for natural and adversarial background variations (Figure 2). Our automated approach generates datasets under varying distribution shifts, useful for benchmarking vision and vision-language models.

In Section 3.1, we outline the preliminaries of the foundational models used. In Section 3.2, we detail our method.

3.1 Preliminaries

Diffusion Models. Diffusion models have significantly advanced in generating high-quality images and refining them based on textual guidance. During training, noisy versions $\mathcal{I}_t$ of the clean image $\mathcal{I}$ are input to the model ϵ_θ at various time steps t, with the goal of learning the noise added at each step. Training consists of two stages: in the forward process *(first stage)*, Gaussian noise from a normal distribution $\mathcal{N}(0, I)$ is incrementally added to $\mathcal{I}$ according to a variance schedule $(\beta_t : t = 1, ..., T)$. Using reparameterization, the noisy image at any time step is:

$$\mathcal{I}_t = \sqrt{\bar{\alpha}_t}\mathcal{I} + \sqrt{1 - \bar{\alpha}_t}\epsilon \quad \epsilon \sim \mathcal{N}(0, I) \tag{1}$$

Here, $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. As $T \to \infty$, $\bar{\alpha}_T \to 0$, meaning $\mathcal{I}_T \sim \mathcal{N}(0, I)$ and all information from $\mathcal{I}$ is lost. Diffusion models are typically conditioned on t, class label $\boldsymbol{y}$, or textual description $\mathcal{T}$, with recent extensions incorporating image $\mathcal{I}$ and its mask for image editing tasks [43,45].

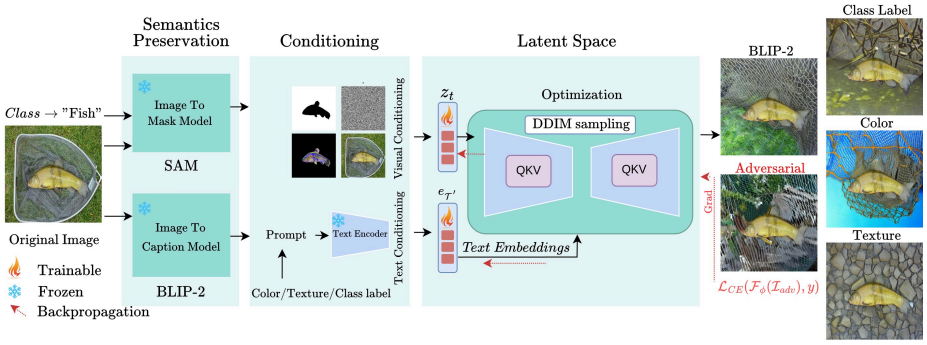

Fig. 2: ObjectCompose uses an inpainting-based diffusion model to generate counterfactual backgrounds. The object mask is obtained from SAM using the class label as a prompt. The segmentation mask and original image caption (from BLIP-2) are fed into the diffusion model. For adversarial examples, both the latent and conditional embeddings are optimized during denoising.

In the reverse process *(second stage)*, the model ϵ_θ learns to approximate the Gaussian parameters at each time step for the reverse conditional distribution. The objective L_t minimizes the error between the predicted and actual noise at each time step:

$$L_t = ||\epsilon - \epsilon_\theta^t(\mathcal{I}_t, e_\mathcal{T}, \psi)||^2 \tag{2}$$

where $e_\mathcal{T}$ is the embedding of the conditional guidance, and ψ represents any additional conditioning, such as masks or scene layouts.

Foundational Models. BLIP-2 [31] introduces an efficient vision-language pre-training approach using a lightweight Querying Transformer (QFormer) to bridge the gap between pre-trained vision and large language models (LLMs). Images are processed by a vision encoder, with relevant features extracted via QFormer and passed to the LLM to generate descriptive captions.

Recently, Segment Anything Model (SAM) [29], an image segmentation model, was introduced that undergoes pre-training on an extensive dataset of high-quality images. SAM uses prompts—such as points, boxes, masks, or text—to identify objects in images. The image is encoded by a transformer-based encoder, and the extracted features, combined with prompt embeddings, are processed by a lightweight decoder to produce the segmentation mask.

3.2 Object-to-Background Compositional Changes

In order to generate object-to-background compositional changes without altering object semantics our method consists of an Object-to-Background Conditioning module to provide strong visual guidance to the text-to-image diffusion model. In the next stage, we condition the diffusion model on the textual prompt

to introduce desired background changes or optimize the latent representation and textual embedding in order to generate adversarial backgrounds.

Preserving Object Semantics. We propose an Object-to-Background Conditioning Module denoted as $\mathcal{C}$, which takes the input image $\mathcal{I}$ and the provided label $\boldsymbol{y}$ as inputs, and returns both the textual prompt $\mathcal{T}$ describing the scene and mask $\mathcal{M}$ encapsulates the object in the image:

$$\mathcal{C}(\mathcal{I}, y) = \mathcal{T}, \mathcal{M} \tag{3}$$

Our conditioning module leverages a promptable segmentation model called SAM [29] denoted by $\mathcal{S}$. By passing the class information $\boldsymbol{y}$ and the image $\mathcal{I}$ to the model $\mathcal{S}(\mathcal{I}, \boldsymbol{y})$, we obtain the object mask $\mathcal{M}$. Simultaneously, to acquire a description for the image scene, we utilize BLIP-2 [32], an image-to-text model denoted as $\mathcal{B}$ to get the necessary prompt $\mathcal{T}_{\mathcal{B}}$ describing the scene, thereby providing object-to-background context information.

$$\mathcal{B}(\mathcal{I}) = \mathcal{T}_{\mathcal{B}} \quad ; \quad \mathcal{S}(\mathcal{I}, \boldsymbol{y}) = \mathcal{M} \tag{4}$$

The mask $\mathcal{M}$ and the textual prompt $\mathcal{T}_{\mathcal{B}}$ serve as conditioning inputs for the subsequent stage, where we employ a diffusion model to generate diverse background variations. This methodical integration of segmentation and language comprehension offers fine-grained control over image backgrounds while upholding object semantics, leading to refined object-centric image manipulations. It's worth noting that we have the flexibility to choose any desired textual prompt $\mathcal{T}$, and are not confined to using $\mathcal{T}_{\mathcal{B}}$ as the textual condition.

Background Generation. Once we've obtained both visual and textual information $(\mathcal{T}, \mathcal{M})$ from our conditioning module, we employ a diffusion model that has been trained for inpainting tasks, which has additional conditioning ψ comprising of the image $\mathcal{I}$ and its corresponding mask $\mathcal{M}$. The denoising operation takes place in the latent space instead of the image pixel space, which is facilitated through the use of a variational autoencoder that provides the mapping between images and their respective latent representations. During the denoising stage, starting with a standard normal Gaussian noise latent z_t, the diffusion model calculates the estimated noise $\hat{\epsilon}^t_\theta$ to be removed from the latent at time step t using a linear combination of the noise estimate conditioned on the textual description $\epsilon^t_\theta(z_t, e_\mathcal{T}, i, m)$ and the unconditioned estimate $\epsilon^t_\theta(z_t, i, m)$:

$$\hat{\epsilon}^t_\theta(z_t, e_\mathcal{T}, i, m) = \epsilon^t_\theta(z_t, i, m) + \lambda \left(\epsilon^t_\theta(z_t, e_\mathcal{T}, i, m) - \epsilon^t_\theta(z_t, i, m)\right) \tag{5}$$

Here, (i, m) represents the representation of the original image $\mathcal{I}$ and its corresponding mask $\mathcal{M}$ in the latent space. The guidance scale λ determines how much the unconditional noise estimate $\epsilon_\theta(z_t, i, m)$ should be adjusted in the direction of the conditional estimate $\epsilon_\theta(z_t, e_\mathcal{T}, i, m)$ to closely align with the provided textual description $\mathcal{T}$(see Appendix A.1). In this whole denoising process, the mask $\mathcal{M}$ generated from our conditioning module guides the image alterations to the background of the object, while the textual description $\mathcal{T}$ contains information for the desired background change.

Our method also handles adversarial background changes by optimizing the conditioned visual and textual latents z_t and $e_{\mathcal{T}}$ through a discriminative model $\mathcal{F}_\phi$ to craft adversaries. For generating adversarial examples the goal of the attacker is to craft perturbations δ that when added to clean image $\mathcal{I}$ with class label $\boldsymbol{y}$, result in an adversarial image $\mathcal{I}_{adv} = \mathcal{I} + \delta$ which elicits an incorrect response from a classifier model $\mathcal{F}_\phi$ i.e., $\mathcal{F}_\phi(\mathcal{I}_{adv}) \neq \boldsymbol{y}$, where ϕ are the model parameters. Usually in pixel-based perturbations, δ is bounded by a norm distance, such as l_2 or l_∞ norm to put a constraint on pixel-level changes done to preserve the semantics of the image. However, in our setting, the control on the amount of perturbation added is governed by the textual and visual latent passed to the diffusion model. In our method (see Algo. **??**), we use the discriminative model $\mathcal{F}_\phi$ to guide the diffusion model ϵ_θ to generate adversarial examples by optimizing its latent representations z_t and $e_{\mathcal{T}}$:

$$\max_{z_t, e_{\mathcal{T}}} \mathcal{L}_{adv} = \mathcal{L}_{CE}(\mathcal{F}_\phi(\mathcal{I}_{adv}), \boldsymbol{y}) \tag{6}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss, $e_{\mathcal{T}}$ is textual embedding and z_t is the denoised latent at time step t. $\mathcal{I}_{adv}$ represents the image generated by the diffusion model after it has been denoised using DDIM [49], a deterministic sampling process in which the latent update is formulated as:

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{z_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}_\theta^t}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1-\bar{\alpha}_{t-1}}\hat{\epsilon}_\theta^t, \quad t = T, \ldots, t-1, \ldots, 1 \tag{7}$$

Our proposed unconstrained adversarial objective $\mathcal{L}_{adv}$ would lead to unrestricted changes in the image background while object semantics are preserved by using the mask conditioning from $\mathcal{S}$.

4 Experimental Protocols

Dataset Preparation. For classification, we initially gathered 30k images from the ImageNet validation set [10], which are correctly classified with high success rate using an ensemble of models; ViT-T, ViT-S [11], Res-50, Res-152 [18], Dense-161 [25], Swin-T, and Swin-S [35]. In order to create a high-quality dataset for our object-to-context variation task, we remove image samples where the boundary between foreground and background is not distinct, e.g., "mountain tent" where the mountain might appear in the background of the tent. This processing results in 15k images. Then for foreground semantic preservation, we utilize a compute-efficient variant of SAM, known as FastSAM [56] with class labels as prompts to generate segmentation masks of the foreground object. However, FastSAM encounter challenges in accurately segmenting objects in all images. To address this, we selected images where the mask-creation process demonstrated exceptional accuracy and generated a clear separation between the object of interest and its background. This meticulous selection process yield a curated dataset comprising 5,505 images, representing a subset of 582 ImageNet classes.

We refer to this dataset as `ImageNet-B`. Due to the computational cost involved in adversarial background optimization and running baseline methods, we select a subset of 1000 images from 500 classes of `ImageNet-B` by sampling two images from each class for comparison. We refer to this dataset as $\texttt{ImageNet-B}_{1000}$. Rest of our experiments are performed on the full `ImageNet-B` dataset.

For object detection, we manually filtered 1,127 images from the COCO 2017 validation set [34], ensuring a clear distinction between foreground objects and background, referred to as `COCO-DC`. This dataset, containing multiple objects per image, is used for both detection and classification. For classification, models are trained on the COCO train dataset using the label of the object with the largest mask region and evaluated on our generated dataset. Additional details and dataset comparisons are provided in Appendix A.16.

Diffusion Parameters. We use the pre-trained Inpaint Stable Diffusion v2 [43] as our text-to-image model and set the guidance parameter λ to 7.5, and use the DDIM sampling [49] with $T = 20$ timesteps. We craft adversarial examples on $\texttt{ImageNet-B}_{1000}$ using Res-50 [18] as the classifier model and maximize the adversarial loss $\mathcal{L}_{adv}$ shown in Eq.6 for 30 iterations. For `COCO-DC`, we maximize the loss in the feature space of the model. Both the text embedding $e_{\mathcal{T}}$ of the prompt $\mathcal{T}$(initialized with $\mathcal{T}_{\mathcal{B}}$) and denoised latent z_t are optimized from denoising time step $t = 4$ using AdamW [36] with a learning rate of 0.1.

Vision Models. We conducted evaluations for the classification task using a diverse set of models. ***a)*** *Natural ImageNet Training:* We evaluate seven naturally ImageNet-trained vision transformers and convolutional neural networks (CNNs). Specifically we use ViT-T, ViT-S [11], Res-50, Res-152 [18], Dense-161 [25], Swin-T, and Swin-S [35]. ***b)*** *Adversarial ImageNet Training:* We also evaluate adversarial ImageNet-trained models including ResAdv-18, ResAdv-50, and WideResAdv-50 at various perturbation budget of ℓ_∞ and ℓ_2 [46]. ***c)*** *Multimodal Training:* Additionally, we explored seven vision language foundational models within CLIP [42] and EVA-CLIP [50]. ***d)*** *Stylized ImageNet Training:* We evaluate the DeiT-T and DeiT-S models trained on a stylized version of the ImageNet dataset [14,39]. ***e)*** *Self-Supervised Training:* We evaluate the performance of Dinov2 models with registers [9,40] which are trained in a self-supervised manner on a large-scale curated dataset LVD-142M, and subsequently fine-tuned on ImageNet. ***f)*** *Segmentation and Detection:* We evaluate Mask-RCNN for segmentation and object detection respectively using our proposed background-to-object variations. Evaluations on FastSAM [56] and DETR [4] are reported in Appendix A.11 and A.10. ***g)*** *Image Captioning:* We also evaluate the robustness of a recent image captioning model BLIP-2 [32], using our generated dataset. For the task ***a)***, and ***b)***, we provide comparison with the baseline methods on $\texttt{ImageNet-B}_{1000}$ and report results on `ImageNet-B` in the Appendix A.5.

Evaluation Metrics: We use the top-1 accuracy (%), Intersection Over Union (IoU), Average Precision(AP) and Recall(AR), and CLIP text similarity score for classification, segmentation, object detection, and captioning tasks, respectively.

Background Conditioning. To induce background variations, we use the following text prompt templates: Class Label: "*A picture of a class*" where *class*

Table 1: Resilience evaluation of vision models on $\texttt{ImageNet-B}_{1000}$(Top-1 (%) accuracy). Our natural object-to-background changes, including color and texture, perform favorably against state-of-the-art methods. Furthermore, our adversarial object-to-background changes show a significant drop in performance across vision models.

Datasets	ViT				CNN			
	ViT-T	ViT-S	Swin-T	Swin-S	Res-50	Res-152	Dense-161	Average
Original	95.5	97.5	97.9	98.3	98.5	99.1	97.2	97.71
ImageNet-E (λ=-20)	91.3	94.5	96.5	97.7	96.0	97.6	95.4	95.50 (-2.21)
ImageNet-E (λ=20)	90.4	94.5	95.9	97.4	95.4	97.4	95.0	95.19 (-2.52)
ImageNet-E (λ_{adv} = 20)	82.8	88.8	90.7	92.8	91.6	94.2	90.4	90.21 (-7.50)
LANCE	80.0	83.8	87.6	87.7	86.1	87.4	85.1	85.38 (-12.33)
Class label	90.5	94.0	95.1	95.4	96.7	96.5	94.7	94.70 (-3.01)
BLIP-2 Caption	85.5	89.1	91.9	92.1	93.9	94.5	90.6	91.08 (-6.63)
Color	67.1	83.8	85.8	86.1	88.2	91.7	80.9	**83.37** (-14.34)
Texture	64.7	80.4	84.1	85.8	85.5	90.1	80.3	**81.55** (-16.16)
Adversarial	18.4	32.1	25.0	31.7	2.0	14.0	28.0	**21.65** (-76.06)

is the image's class name; Caption: "*captions from BLIP-2*"; Color: "*A picture of* ___ *background*" where ___ is red, green, blue, or colorful; Texture: "*A picture of* __ *background*" with ___ replaced by textured, rich textures, colorful textures, distorted textures; Adversarial: "*captions from BLIP-2*" with updated prompts after optimization. For ImageNet-E [33], default values of λ are employed to regulate the strength of texture complexity. For LANCE [41], we use the default prompt to generate background variations via a large language model. We report the worst-performing prompt across colors and textures, with detailed analysis in Appendix A.5.

4.1 Comparison with Baseline Methods

Natural ImageNet Training. In Table 1, we observe that background variations introduced by our method are more challenging for vision models, resulting in a performance drop of 13.5% compared to ImageNet-E ($\lambda = 20$) on natural background variations. When subjected to adversarial background changes, a substantial performance drop of 68.56% is observed compared to ImageNet-E ($\lambda_{adv} = 20$), highlighting the effectiveness of the unconstrained nature of our attack. Background variations by our method show a consistent decline in accuracy for both transformer-based and CNN models when exposed to diverse object-to-background changes. This decrease is especially noticeable in texture and color backgrounds. We find that as we moved from purely transformer-based architectures to convolution-based architectures, there is an overall improvement in accuracy across natural background changes. For instance, the average accuracy across all backgrounds for ViT-T, Swin-T, and Res-50 on $\texttt{ImageNet-B}_{1000}$ is 76.95%, 89.22% and 91.08% respectively. Further, we observe that as the model capacity is increased across different model families, the robustness to background changes also increases. As is evident, the models are most vulnerable to

Table 2: Resilience evaluation of Zero-shot CLIP and EVA-CLIP models on $\texttt{ImageNet-B}_{1000}$(Top-1 (%) accuracy). Our natural object-to-background changes, including color and texture, perform favorably against state-of-the-art methods. We find that EVA-CLIP models show better performance across all background variations.

Datasets	CLIP							
	ViT-B/32	ViT-B/16	ViT-L/14	Res50	Res101	Res50x4	Res50x16	Average
Original	73.90	79.40	87.79	70.69	71.80	76.29	82.19	77.43
ImageNet-E (λ=-20)	69.79	76.70	82.89	67.80	69.99	72.70	77.00	73.83 (-3.60)
ImageNet-E (λ=20)	67.97	76.16	82.12	67.37	39.89	72.62	77.07	73.31 (-4.12)
ImageNet-E (λ_{adv} = 20)	62.82	70.50	77.57	59.98	65.85	67.07	67.07	68.23 (-9.20)
LANCE	54.99	54.19	57.48	58.05	60.02	60.39	73.37	59.78 (-17.65)
Class label	78.49	83.69	88.79	76.60	77.00	82.09	84.50	81.59 (+4.16)
BLIP-2 Captions	68.79	72.29	79.19	65.20	68.40	71.20	75.40	71.49 (-5.94)
Color	48.30	61.00	69.51	50.50	54.80	60.30	69.28	**59.14** (-18.29)
Texture	49.60	62.39	66.99	51.69	53.20	60.79	67.49	**58.88** (-18.55)
Adversarial	25.5	34.89	48.19	18.29	24.40	30.29	48.49	**32.87** (-46.25)
Datasets	EVA-CLIP							
	g/14	g/14+	B/16	L/14	L/14+	E/14	E/14+	Average
Original	88.80	92.69	89.19	91.10	91.99	93.80	94.60	91.74
ImageNet-E (λ=-20)	84.74	88.98	85.55	89.19	88.78	92.02	91.81	88.72 (-3.02)
ImageNet-E (λ=20)	84.10	89.40	85.81	88.51	89.69	92.69	92.50	88.95 (-2.79)
ImageNet-E (λ_{adv} = 20)	79.69	85.45	80.20	84.04	85.95	89.89	89.59	84.97 (-6.77)
LANCE	70.25	77.40	73.26	76.63	77.46	80.95	78.65	76.37 (-15.37)
Class label	90.10	92.90	88.61	91.31	91.90	93.40	93.41	91.66 (-0.08)
BLIP-2 Caption	80.31	84.29	82.10	82.50	84.80	86.90	86.90	83.97 (-7.77)
Color	73.50	80.50	73.20	80.70	84.61	84.39	87.00	80.55 (-11.19)
Texture	75.30	78.90	74.40	80.80	82.10	83.60	85.60	80.10 (-11.64)
Adversarial	55.59	62.49	48.70	65.39	73.59	70.29	73.29	**64.19** (-27.55)

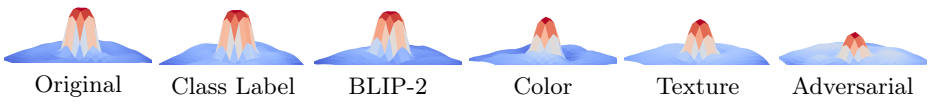

Fig. 3: The loss surfaces *(flipped)* of the ViT-S depicted on `ImageNet-B`. Significant distribution shifts result in narrow and shallow surfaces at convergence.

adversarial background changes, resulting in a significant drop in average accuracy. Res-50 shows most drop on adversarial changes, which is expected as it serves as the discriminative model $\mathcal{F}_\phi$ (Eq. 6) for generating adversarial examples. In Figure 3, we depict the loss surfaces of ViT-S and observe that these surfaces become narrower and shallower with more pronounced background variations, aligning with our results. We provide results on `ImageNet-B` and `COCO-DC` dataset in Appendix A.5 and A.7 with ablations across different background prompts. Visualizations are provided in Figure 4 and Appendix A.13.

Multimodal Training. In Table 2, we observe that compared to ImageNet-E (λ = 20), our natural background variations lead to an average performance drop of 15.66% and 8.85% on CLIP and EVA-CLIP models. On comparing with ImageNet-E (λ_{adv} = 20), our adversarial background variations lead to an average performance drop of 35.36% and 20.78% on CLIP and EVA-CLIP mod-

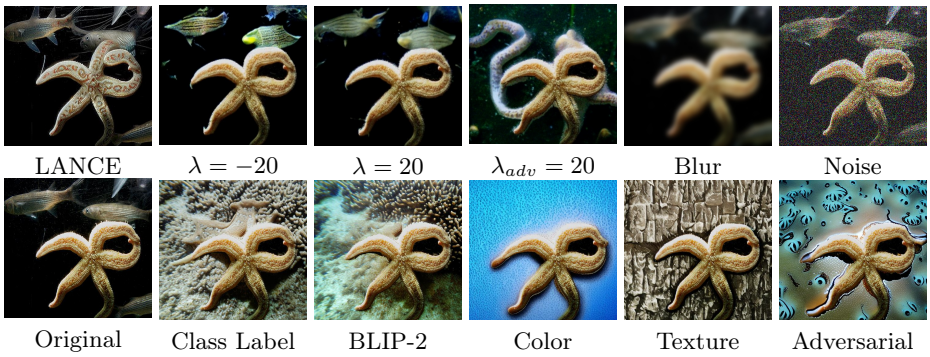

Fig. 4: Qualitative comparison of our method *(bottom row)* with previous related work *(top row)*. Our method enables diversity and controlled background edits.

els. Similar to results mentioned in Table 1, zero-shot robustness shows similar trend across different background changes. However, for background variations induced using class label information the performance increases in CLIP-based models. This reason could be the use of CLIP text encoder utilized for generating the textual embedding $e_{\mathcal{T}}$ for guiding the generation process of the diffusion model. On EVA-CLIP, which proposed changes to stabilize the training of CLIP models on large-scale datasets, we observe significant improvement in zero-shot performance across all background changes. In Appendix A.5, we delve into the comparison between multimodal and unimodal models, offering detailed results on `ImageNet-B` dataset.

Object Semantic distortion: It's noteworthy to mention that in both Table 1 and 2, we observe a significant drop in performance of models across background changes induced by LANCE method. However, we discover that the drop in performance is not necessarily due to the induced background changes, rather than distorting the object semantics, making it unsuitable for evaluating object-to-background context. This observation is supported by Figure 5, where we evaluate performance on original and LANCE generate images while masking the background. A significant performance drop is evident across all models, emphasizing the distortion of object semantics. We provide a detailed discussion with FID [23] comparison and visualizations in Appendix A.3 and A.4.

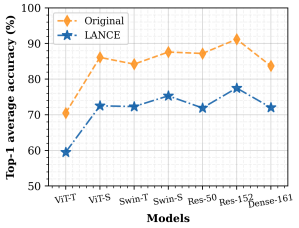

Fig. 5: Evaluating LANCE on $\texttt{ImageNet-B}_{1000}$ dataset with masked background.

4.2 Further Evaluations

Adversarial ImageNet Training. As can be seen from Figure 6 *(bottom row)*, our object-to-background compositional changes on `ImageNet-B` lead to a sig-

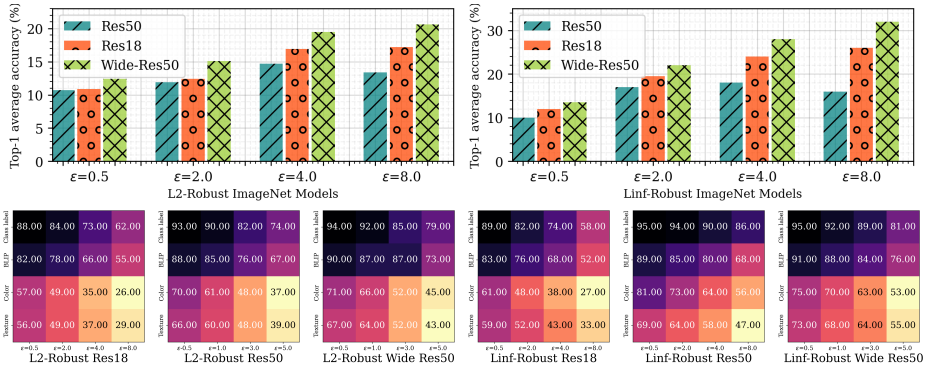

Fig. 6: The top row plots the Top-1(%) accuracy achieved by adversarially trained ResNet models on adversarial background changes on $\texttt{ImageNet-B}_{1000}$ and the bottom row indicates for the case of non-adversarial background changes on `ImageNet-B`.

nificant decline in accuracy for adversarially trained models. This highlights the robustness of these models is limited to adversarial perturbations and does not transfer to different distribution shifts. In Figure 6 *(top row)*, when we evaluate these models on adversarial background changes on $\texttt{ImageNet-B}_{1000}$, the performance improves with an increase in adversarial robustness(ϵ) of the models. Furthermore, we also observe models with more capacity perform better, similar to results on natural training. For detailed results and comparison with baseline methods, refer to Appendix A.6.

Stylized ImageNet Training. Despite the focus of Stylized ImageNet training [14] to encourage models to concentrate on the foreground of the scene by reducing background cues for prediction [39], our findings indicate that it is still susceptible to both natural and adversarial object-to-background variations (see Table 3). Consequently, its applicability appears to be constrained to specific distribution shifts.

Self-Supervised Training. Improved performance is observed in Dinov2 models across object-to-background variations (see Figure 7). We hypothesize this improvement is acquired through training on extensive curated datasets and the utilization of additional learnable registers/tokens during training for refining the interpretability of attention maps. For more details, refer to Appendix A.8.

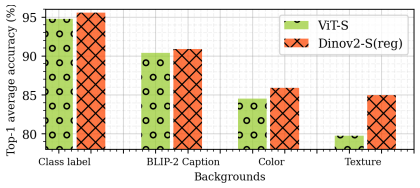

Fig. 7: Evaluating Dinov2 models on `ImageNet-B` background changes.

Segmentation and Detection. We observe a consistent decrease in AP scores on object detection and instance segmentation tasks across background variations generated on `COCO-DC`(see Table 5). The adversarial background results in the lowest AP scores, but still remains at a reasonable level given that the adversarial examples are generated using a classification model, with limited

Table 3: Stylized Training Evaluation

Datasets	Background	Stylized Trained models		
		DeiT-S	DeiT-T	Average
ImageNet-B	Original	91.22	87.21	89.21
	Class label	89.35	85.35	87.35 (-1.86)
	BLIP-2 Caption	84.01	79.19	81.60 (-7.61)
	Color	66.57	57.54	62.05 (-27.15)
	Texture	64.08	54.82	59.45 (-29.76)
ImageNet-B_{1000}	Original	89.60	85.90	87.75
	Adversarial	15.90	10.80	13.35 (-74.40)

Table 4: Image-to-Caption (BLIP-2) Evaluation

Dataset	Background	CLIP Score
ImageNet-B	Class Label	0.75
	BLIP-2 Caption	0.84
	Color	0.66
	Texture	0.67
ImageNet-B_{1000}	Adversarial	0.62

Table 5: Mask AP and Segment AP score on COCO-DC

Background	Box AP	Segment AP
Original	57.99	56.29
BLIP-2 Caption	47.40	44.75
Color	48.12	45.09
Texture	45.79	43.07
Adversarial	37.10	34.91

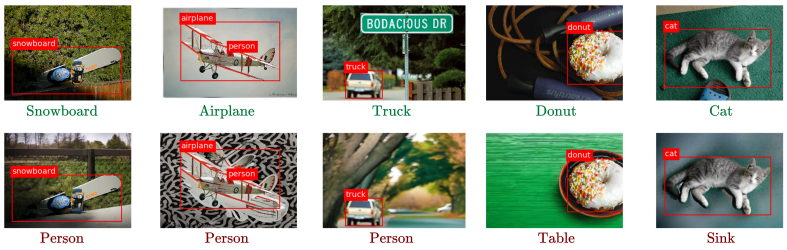

Fig. 8: Correct predictions by Mask-RCNN and Res-50 on the original image *(top row)* and the corresponding predictions on altered backgrounds *(bottom row)*.

cross-task transferability. Moreover, our qualitative observations suggest detection and segmentation models exhibit greater resilience to changes in the background compared to classifiers (see Figure 8 and Appendix A.10.

Image Captioning. Table 4 shows the CLIP scores between captions from clean and generated images using the BLIP-2 model. Scores decrease with color, texture, and adversarial background changes (Appendix A.9 for qualitative results).

5 Conclusion

In this study, we propose ObjectCompose, a method for generating object-to-background compositional changes. Our method addresses the limitations of current works, specifically distortion of object semantics and diversity in background changes. We accomplish this by utilizing the capabilities of image-to-text and image-to-segmentation foundational models to preserve the object semantics, while we optimize for diverse object-to-background compositional changes by modifying the textual prompts or optimizing the latents of the text-to-image model. ObjectCompose offers a complimentary evaluation protocol to the existing ones, for comprehensive evaluations across current vision-based models to reveal their vulnerability to background alterations. In Appendix A.18, we elaborate on the initial insights gained from our work and discuss current limitations and future directions.

References

1. Alcorn, M.A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.S., Nguyen, A.: Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4845–4854 (2019)
2. Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., Katz, B.: Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. Advances in neural information processing systems **32** (2019)
3. Bordes, F., Shekhar, S., Ibrahim, M., Bouchacourt, D., Vincent, P., Morcos, A.S.: Pug: Photorealistic and semantically controllable synthetic data for representation learning (2023)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
5. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
6. Chen, J., Chen, H., Chen, K., Zhang, Y., Zou, Z., Shi, Z.: Diffusion models for imperceptible and transferable adversarial attack (2023)
7. Christensen, P.E., Snæbjarnarson, V., Dittadi, A., Belongie, S., Benaim, S.: Assessing neural network robustness via adversarial pivotal tuning. arXiv preprint arXiv:2211.09782 (2022)
8. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501 (2018)
9. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Fu, Y., Zhang, S., Wu, S., Wan, C., Lin, Y.: Patch-fool: Are vision transformers always robust against adversarial perturbations? arXiv preprint arXiv:2203.08392 (2022)
13. Gao, S., Li, Z.Y., Yang, M.H., Cheng, M.M., Han, J., Torr, P.: Large-scale unsupervised semantic segmentation. IEEE transactions on pattern analysis and machine intelligence (2022)
14. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018)
15. Gondal, M.W., Wuthrich, M., Miladinovic, D., Locatello, F., Breidt, M., Volchkov, V., Akpo, J., Bachem, O., Schölkopf, B., Bauer, S.: On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. Advances in Neural Information Processing Systems **32** (2019)
16. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)

17. Haralick, R.M., Shanmugam, K., Dinstein, I.H.: Textural features for image classification. IEEE Transactions on systems, man, and cybernetics (6), 610–621 (1973)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8349 (2021)
20. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019)
21. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15262–15271 (2021)
22. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
23. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
24. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
25. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
26. Idrissi, B.Y., Bouchacourt, D., Balestriero, R., Evtimov, I., Hazirbas, C., Ballas, N., Vincent, P., Drozdzal, M., Lopez-Paz, D., Ibrahim, M.: Imagenet-x: Understanding model mistakes with factor of variation annotations. arXiv preprint arXiv:2211.01866 (2022)
27. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2901–2910 (2017)
28. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113–19122 (2023)
29. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
30. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236 (2016)
31. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
32. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023)
33. Li, X., Chen, Y., Zhu, Y., Wang, S., Zhang, R., Xue, H.: Imagenet-e: Benchmarking neural network robustness via attribute editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20371–20381 (2023)

34. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2015)
35. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows (2021)
36. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
37. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
38. Moayeri, M., Pope, P., Balaji, Y., Feizi, S.: A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes (2022)
39. Naseer, M., Ranasinghe, K., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Intriguing properties of vision transformers (2021)
40. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
41. Prabhu, V., Yenamandra, S., Chattopadhyay, P., Hoffman, J.: Lance: Stress-testing visual models by generating language-guided counterfactual images (2023)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
43. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
44. Rosenfeld, A., Zemel, R., Tsotsos, J.K.: The elephant in the room. arXiv preprint arXiv:1808.03305 (2018)
45. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
46. Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., Madry, A.: Do adversarially robust imagenet models transfer better? (2020)
47. Sharma, A., Bian, Y., Munz, P., Narayan, A.: Adversarial patch attacks and defences in vision-based tasks: A survey. arXiv preprint arXiv:2206.08304 (2022)
48. Shetty, R., Schiele, B., Fritz, M.: Not using the car to see the sidewalk–quantifying and controlling the effects of context in classification and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8218–8226 (2019)
49. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
50. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023)
51. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
52. Xiao, K., Engstrom, L., Ilyas, A., Madry, A.: Noise or signal: The role of image backgrounds in object recognition (2020)
53. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019)

54. Zhang, C., Pan, F., Kim, J., Kweon, I.S., Mao, C.: Imagenet-d: Benchmarking neural network robustness on diffusion synthetic object. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21752–21762 (2024)
55. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
56. Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., Wang, J.: Fast segment anything (2023)
57. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
58. Zhu, Z., Xie, L., Yuille, A.L.: Object recognition with and without objects. arXiv preprint arXiv:1611.06596 (2016)