XIMAGENET-12: An Explainable Visual Benchmark Dataset for Model Robustness Evaluation

Qiang Li[1,*]
RWTH Aachen, Germany
qiang.li@rwth-aachen.de

Dan Zhang[1]
SensLab Technology, China
dannie2023.zhang@gmail.com

Shengzhao Lei [2]
EPFL, Switzerland
shengzhaolei@gmail.com

Xun Zhao[2]
University of Amsterdam, Netherlands
x.zhao@uva.nl

Weiwei Li[3]
Shanghai Business School, China
23154040112@stu.sbs.edu.cn

Porawit Kamnoedboon[3]
University of Zurich, Switzerland
porawit.kamnoedboon@uzh.ch

Junhao Dong [3]
Nanyang Technological University, Singapore
junhao003@ntu.edu.sg

Shuyan Li *
University of Cambridge, UK
sl2141@cam.ac.uk

Abstract

Despite the promising performance of existing visual models on public benchmarks, the critical assessment of their robustness for real-world applications remains an ongoing challenge. To bridge this gap, we propose an explainable visual dataset, XIMAGENET-12, to evaluate the robustness of visual models. XIMAGENET-12 consists of over 200K images with 15,410 manual semantic annotations. Specifically, we deliberately selected 12 categories from ImageNet, representing objects commonly encountered in practical life. To simulate real-world situations, we incorporated six diverse scenarios, such as overexposure, blurring, and color changes, etc. We further develop a quantitative criterion for robustness assessment, allowing for a nuanced understanding of how visual models perform under varying conditions, notably in relation to the background. We make the XIMAGENET-12 dataset and its corresponding code openly accessible at https://sites.google.com/view/ximagenet-12/home. *We expect the introduction of the XIMAGENET-12 dataset will empower researchers to thoroughly evaluate the robustness of their visual models under challenging conditions.*

1. Introduction

Visual models have been widely utilized in a variety of real-world applications, including manufacturing, maintenance, etc. [10, 20, 23, 44, 47]. Despite their commendable performance on standardized benchmark datasets, existing visual models often exhibit noticeable performance degradation in real-world deployments [3, 14, 20, 41]. Challenges such as variations in lighting, background interference, object displacements and unexpected environmental factors, like noises or artificial camera disturbances, are common issues encountered by visual models in practical scenarios [23, 42, 44].

The lack of a readily available and interpretable dataset makes the evaluation of robustness an open challenge. There are a few works attempting to explore how existing visual models are influenced by contextual bias or backgrounds [27, 28, 30, 37, 43, 48]. Among them, the most similar work is ImageNet-9 dataset [37], which selects nine classes and explores the impact of backgrounds on foreground objects. However, their work did not deeply investigate what factors in the backgrounds really matter for the model behavior, thus leading to less explainability. Besides, their semantic labeling is not precise enough and such rough segmentation of foreground and background leads to some misleading conclusions.

In this work, we propose an explainable visual benchmark dataset, XIMAGENET-12, to evaluate the robustness

[1] indicates Co-first Authorship, and * indicates Shared Corresponding Author. Acknowledging the great support and sponsorship received from Accenture, V7 Lab and Playground AI. Paper accepted by Synthetic Data for Computer Vision Workshop @ IEEE CVPR 2024

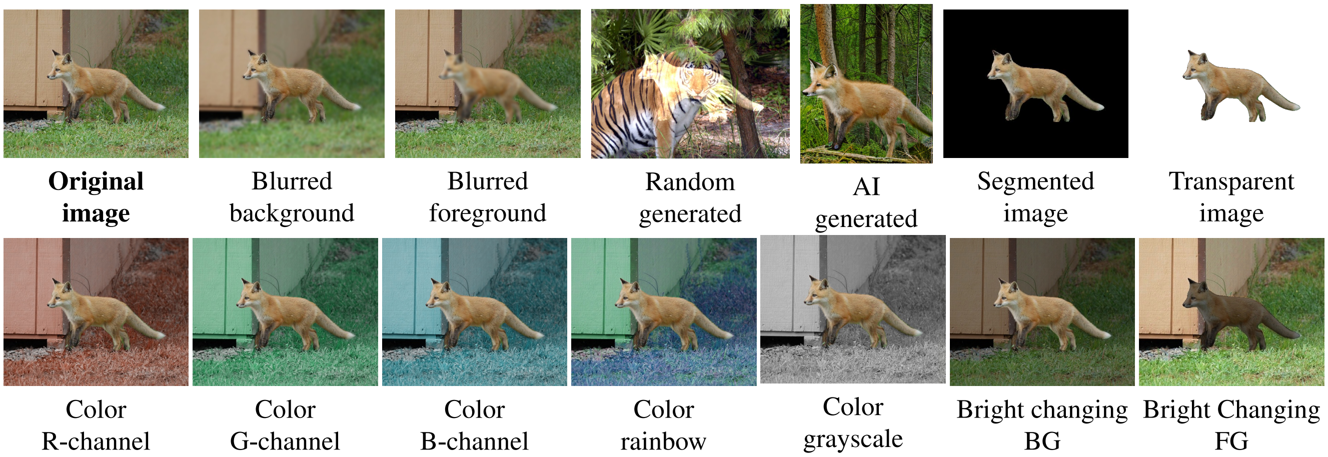


Figure 1. XIMAGENET-12 sample for 6 scenarios: Blur, Random generated background, AI-generated background, Segmentated, Transparent and Color images. Over 200K Images in total.

of visual models when facing challenging real-world scenarios. XIMAGENET-12 consists of over 200K images with 15,410 manual semantic annotations. Specifically, we deliberately selected 12 categories from ImageNet [7]. These images contain objects that are commonly found in daily life, with relatively complicated backgrounds. To simulate real-world situations, we incorporated six diverse scenarios that often occur in real-world applications. As shown in Figure 1, these scenarios cover background & foreground blurring, color changes to simulate the camera vibrations in industrial production processes, as well as background replacement & removal and artificially rendered backgrounds for enhanced validation. It is worth noting that our semantic annotations of foreground and background are precise, which allows us to deeply investigate how the visual model is influenced by the backgrounds. We further develop a quantitative criterion for robustness assessment, allowing for a comparative evaluation of visual model robustness. We show that the robustness score of visual backbones calculated on our dataset can provide guidance for practical visual model usage. We summarize our main contributions as follows:

- We create a dataset, named XIMAGENET-12, consisting of a variety of challenging scenarios.
- We develop a quantitative criterion to evaluate the robustness of visual models and show its effectiveness in providing guidance for real-world applications.
- We deeply investigate the influence of backgrounds and show some interesting findings based on our well-annotated dataset.

2. Related Work

In this section, we discuss previous works that investigate models' performances dependence with contextual bias and backgrounds. Previous research has studied the overarching phenomenon of contextual bias [16, 31, 35], proposing methods to mitigate its impact. For example, Khosla *et al.* proposed a discriminative framework that directly exploited dataset bias during training [16]. Torralba *et al.*compared multiple popular datasets by using a variety of evaluation criteria to obtain directions that could improve dataset collection and algorithm evaluation protocols [35].

Among them, the works most similar to ours are proposed by Zhu *et al.* [48] and Xiao *et al.* [37], both of which delved into ImageNet[7] classification and segmentation, and background exploration. Zhu *et al.* [48] trained deep neural networks on the foreground and background respectively, demonstrating that valuable visual hints can be learned separately and then combined to achieve higher performance. As Zhu *et al.* [48] did not conduct the evaluation of recent visual models, Xiao *et al.* [37] only plainly investigated the influence of backgrounds with state-of-the-art (SOTA) visual models: Noise or Signal. Meanwhile, they proposed a synthetic dataset, ImageNet-9 by disentangling foreground and background signals on ImageNet. Compared with ImageNet-9 [37], our dataset has more precise semantic labels and we demonstrate that poor semantic label quality can yield sub-optimal results through extensive experiments. Furthermore, our dataset encompasses six scenarios simulating challenges commonly encountered in real-world applications.

3. XIMAGENET-12 Dataset

3.1. Dataset Simulation

The overall dataset generation flow is shown in Figure 2. We select 12 categories of images from the ImageNet [7] dataset as the base images. These 12 categories are: **lizard, green lizard, crayfish, whale, dog, fox, anemone fish,**

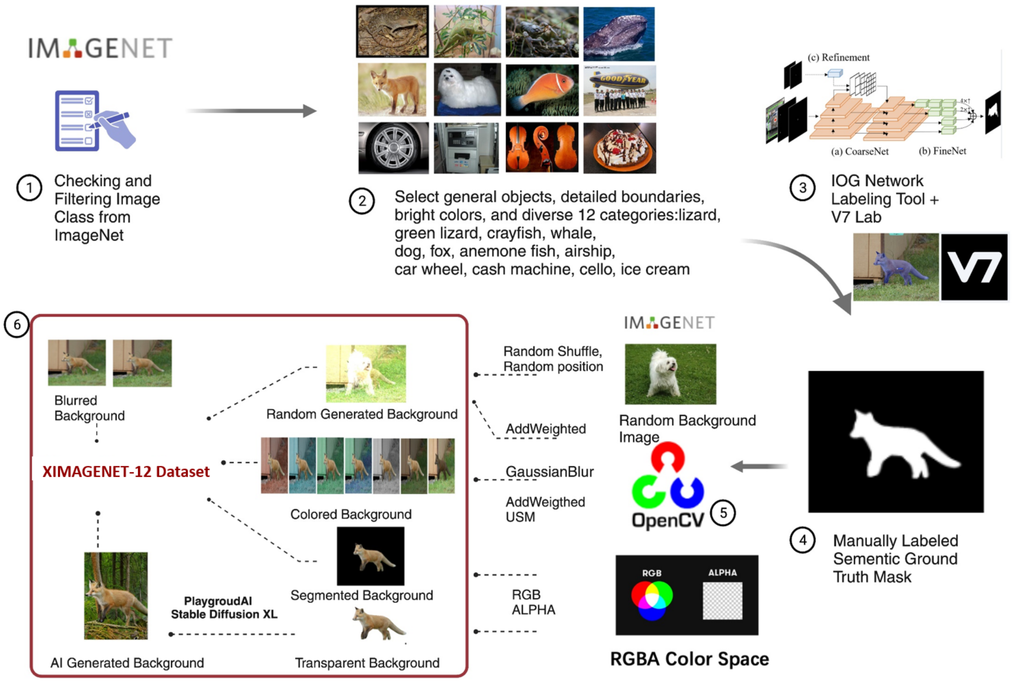

IMAGENET
1 Checking and Filtering Image Class from ImageNet
2 Select general objects, detailed boundaries, bright colors, and diverse 12 categories:lizard, green lizard, crayfish, whale, dog, fox, anemone fish, airship, car wheel, cash machine, cello, ice cream
(c) Refinement
(a) CoarseNet
(b) FineNet
3 IOG Network Labeling Tool + V7 Lab
V7
6
Blurred Background
Random Generated Background
XIMAGENET-12 Dataset
Colored Background
Segmented Background
PlaygroudAI Stable Diffusion XL
AI Generated Background
Transparent Background
Random Shuffle, Random position
AddWeighted
GaussianBlur
AddWeigthed USM
RGB ALPHA
IMAGENET
Random Background Image
OpenCV 5
RGB
ALPHA
RGBA Color Space
4 Manually Labeled Sementic Ground Truth Mask

Figure 2. The flow chart of XIMAGENET-12 generation.

airship, car wheel, cash machine, cello, and ice cream. These selected images contain objects with complex shapes, detailed edges or boundaries, distinguished colors, diverse resolutions, and complex backgrounds, which can represent situations frequently encountered in daily life. To separate the background and foreground, we employ the IOG network [45] for rough segmentation and manually refine the annotations via V7 Lab [36]. Then we synthesize 6 scenarios, including colored images, blurred images, segmented images, transparent images, images with randomly generated backgrounds, and images with AI-generated backgrounds. We detail each scenario as follows.

Colored images: Colored images can simulate lighting changes in the real world. There are 7 different transformations with regards to colored images, including transforming backgrounds to grayscale, single-channel (R, G, B), rainbow, and switching brightness of both backgrounds (bright changing BG) and foregrounds (bright changing FG). Specifically, we use the OpenCV function addWeighted to adjust brightness and sharpness via Unsharp Masking (USM). We generate rainbow images by converting the image to HSV color space and changing the hue of the backgrounds.

Blurred images: Blur often happens when a camera suffers a slight shift, resulting in the degradation of details. We use the OpenCV function GaussianBlur to blur images both in backgrounds (blurred background) and foregrounds (blurred foreground).

Segmented images: We remove the backgrounds of the images and keep the foreground only. Specifically, we keep the RGB channel unchanged for the foreground and set the RGB channel of background as (0, 0, 0).

Transparent images: We create a new image with RGBA 4 channels, where the background is completely removed. For example, if the pixel at (x, y) in the original image is (r,g,b), we set it as (0, 0, 0, 0) if it belongs to the background, and (r, g, b, 255) vice reverse.

Images with randomly generated backgrounds: We randomly select an image from ImageNet [7] dataset as the background and blend it with the foreground of the original image by using addWeighted blending, resize, random Shuffle & position functions.

Images with AI-generated backgrounds: We use the transparent images as inputs for Playground AI [25] with the Stable Diffusion XL model [26]. We have tried different text prompts and found some very useful tips: Using keywords such as 'National Geographic Magazine' or 'National Oceanic Magazine' can increase the authenticity of the generated background; Adding specific and appropriate environmental information to the prompt can make the generated background and objects better integrated. By using a diffusion model and introducing unexpected or extreme background variations, we can assess whether the model is resilient against potential adversarial attacks involving background manipulation.

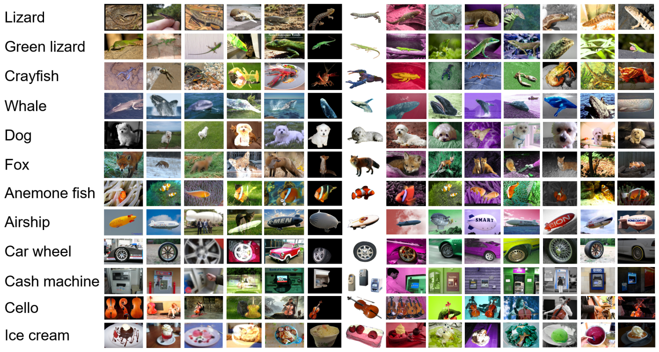


Figure 3. XIMAGENET-12 samples for each class and scenario.

3.2. Dataset Properties

There are 15,410 original images in XIMAGENET-12, with around 1,300 samples in each category. Each image is simulated to 6 scenarios. As we found that some AI-generated images contained too small or incomplete objects, we filtered those sup-optimal images and obtained 12,248 images for AI-generated scenarios, with approximately 1,000 images for each class. Finally, we have in total 212,747 images contained in XIMAGENET-12. Figure 3 illustrates the overall look of our dataset across various scenarios, including original images.

4. Robustness Score

In this section, we introduce our proposed criterion for evaluating the robustness of visual models. We aim to quantitatively assess a model's generalization performance across diverse scenarios. Drawing inspiration from mathematical concepts like variance and covariance, we have devised a robustness score based on the XIMAGENET-12 dataset, as outlined below.

Firstly, we measure the robustness of models in cross scenarios in a variance-like form:

$$\sigma_{cross}^2 = \frac{\sum_{i=1}^{n}(C(i) - \mu)^2}{n}. \tag{1}$$

Here, μ means the best weight accuracy when the model is both trained and tested on the original scenario. $C(i)$ means the model is trained on original images but tested on the i-th scenario. $i \in \{0, 1, \ldots, n\}$ and n is the number of scenarios that we consider.

Similarly, we formulate the robustness of models when trained and tested within the same scenario as follows:

$$\sigma_{inner}^2 = \frac{\sum_{i=1}^{n}(C'(i) - \mu)^2}{n}. \tag{2}$$

Here $C'(i)$ means that the model is both trained and tested on the i-th scenario. Considering both above-mentioned cases, we derive the robustness score as follows:

$$S_{\text{robust}} = 1 - (\sigma_{cross}^2 + \sigma_{inner}^2). \tag{3}$$

We consider a larger robustness score as an indicator of the higher robustness of the visual model.

5. Experiments

5.1. Experimental Settings

In this section, we evaluated the robustness of commonly used visual models with our proposed XIMAGENET-12 dataset and investigated how visual models perform under various conditions. Specifically, we tested classification models with the following selected visual backbones: ResNet [11] series, MobileNet [29], EfficientNet [34] series, InceptionNet [33], DenseNet [13], ViT [8] and Swin Transformers [22]. We included the following segmentation models with the above-mentioned backbones for further evaluation, including PSPNet [46], FPN [21], UperNet [38], DeepLabV3 [4] and DeepLabv3plus R50-D8 [5].

We conducted our experiments by using TensorFlow [1], Keras [15], PyTorch [24], and MMsegmentation Library [6]. For the inputs of classification models, we cropped the images as 224 × 224. For the inputs of segmentation models, we cropped the images as 256 × 256. We trained these models by using the Adam [17] optimizer under the learning rate of 0.0001, with the epochs of 200 and the batch size of 16.

We adopted Top-1 accuracy as the major evaluation metric for classification. We performed Multiple Linear regression [39] to evaluate our hypotheses. We utilized the P-value [9] of 95% CI as the confidence of verified hypotheses. We employed the Variable that accounts for variations across different models, scenarios, and object classes. Estimate in the Table 3 serves as a valuable indicator for assessing accuracy change compared with the reference model. For segmentation models, we used Mean Intersection over Union (MIoU) and accuracy.

5.2. Main Results

Comparison of SOTA Visual Models. Here, we investigate the performance of SOTA models facing diverse scenarios. We study two cases: 1) EX1 Cross Scenario. In this setting, we train the classification model on original images and test them in different scenarios respectively. 2) EX2 Within the same Scenario. In this setting, we both train and test the classification model within the same scenario. We report the classification performance (Top-1 Accuracy) in Table 1. In general, different scenarios influence these models to different degrees. Among those scenarios, removing

Table 1. Comparison of SOTA visual models with diverse scenarios. Here all the evaluation metrics are Top-1 Accuracy.

Pretrained Dataset	**Model Name**	Parameters (M)	**Test Dataset (Top-1 Acc.)**							
			Blur_bg	**Blur_obj**	**Color_g**	**Color_b**	**Color_grey**	**Color_r**	**Rand_bg**	**Seg_img**
ImageNet [7] (Original images) EX1	ResNet50 [11]	25.60	90.97%	88.17%	84.42%	86.98%	92.13%	89.03%	22.41%	68.55%
	VGG-16 [32]	138.4	89.92%	89.91%	78.64%	70.46%	81.48%	80.68%	24.58%	49.62%
	MobileNetV2 [29]	3.5	92.34%	88.52%	85.73%	88.67%	88.81%	89.33%	27.14%	66.43%
	EfficientNetB0 [34]	5.3	91.44%	90.86%	78.10%	82.45%	86.44%	83.65%	25.29%	53.56%
	EfficientNetB3 [34]	12.3	86.80%	84.53%	77.99%	81.22%	83.00%	83.85%	22.06%	69.67%
	DenseNet121 [13]	8.1	93.77%	88.92%	87.39%	87.33%	93.23%	88.21%	26.41%	69.67%
	ViT [8]	86.6	88.44%	90.77%	65.87%	62.82%	70.69%	66.53%	17.21%	49.01%
	Swin [22]	87.76	80.97%	81.57%	64.59%	65.91%	69.28%	64.41%	19.43%	44.57%
XImageNet-12 (*Scenarios) EX2	ResNet50 [11]	25.60	83.52%	80.24%	83.61%	84.45%	84.71%	80.40%	53.91%	85.76%
	VGG-16 [32]	138.4	74.85%	71.54%	74.18%	76.26%	77.58%	69.91%	70.25%	73.27%
	AlexNet [19]	61.1	81.60%	79.95%	81.96%	81.89%	81.31%	78.07%	46.29%	82.00%
	MobileNetV3 [12]	3.50	67.36%	67.88%	72.04%	74.25%	69.48%	64.79%	43.33%	78.85%
	DenseNet121 [13]	8,1	90.79%	86.57%	88.92%	89.96%	90.44%	87.37%	69.58%	91.60%
	ViT [8]	86.56	71.51%	70.21%	74.77%	75.96%	75.80%	71.14%	38.01%	78.69%
	Swin [22]	87.76	72.81%	75.02%	81.05%	81.96%	81.63%	76.42%	13.23%	80.64%

the backgrounds and randomly substituting the background result in most performance drops.

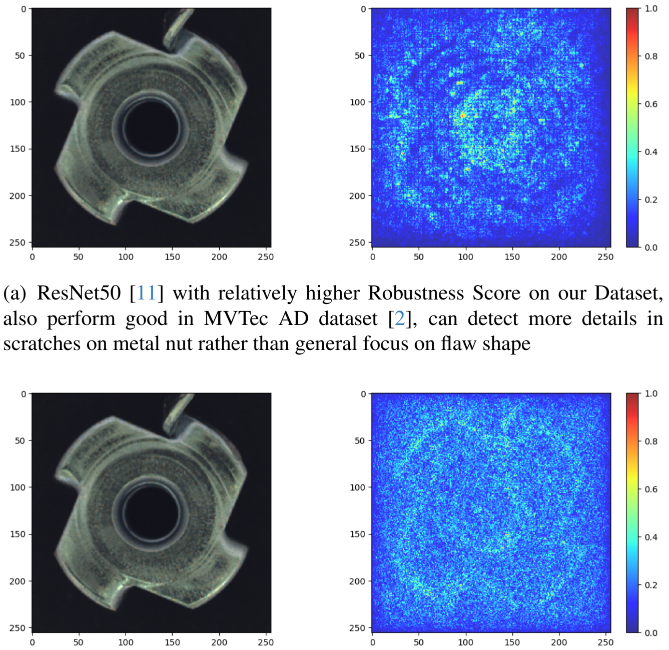
In Table 1 ”Rand_bg” scenario of EX1, all these models show poor performance when trained on original images and tested on images with random backgrounds. This indicates that all these models tend to capture significant information from original backgrounds during training. Those random backgrounds in the test set will heavily interrupt the recognition of visual models. Compared with EX1, the performance drop of EX2 is not so significant. This indicates that when trained on images with random backgrounds, visual models may be aware of the irrationality of the backgrounds and automatically ignore them.

In Table 1 “Seg_img” scenario of EX2, we find that training on the images with removed backgrounds does not lead to a test accuracy drop. This finding contrasts with the assertion by Xiao et al. [37], who claimed that removing the background negatively impacts test accuracy. The suboptimal performance obtained by Xiao et al. is due to the poor annotation quality of ImageNet-9 instead of the missing background (we will further validate this in Sec 5.3). Our experiment on XIMAGENET-12 indicates that ***models trained and tested with well-segmented foregrounds tend to perform well even if the backgrounds are missing***.

By observing blurring and color scenarios in Table 1, CNN-based models generally show better performance with EX1 setting than EX2, while transformer-based models show the opposite results in color scenarios. When observing the Color scenarios (Color_g to Color_r) in EX1, CNN-based models show higher robustness (e.g. the drop rate of VGG-16 [32] is 11.28%) than transformers (the drop rate of ViT [8] is 27.95%).

While ViT [8] and Swin-Transformers [22] show good performance in most visual tasks [40], their accuracy and robustness are not always as good as CNN-based models when facing challenging scenarios. This motivates us with a hypothesis that ***a model with higher accuracy is not necessarily more stable***. We argue that ***more robust models tend to rely less on backgrounds***. To validate our hypothesis, we provide a deeper investigation of the robustness of existing visual models by using our robustness score and statistical analysis in the following parts.

Evaluation of Robustness. Using our proposed robustness score, we quantitatively evaluate the robustness of commonly used visual models and report the results in Table 2. We find that a model with a higher robustness score is more resistant to the changes in background (with lower



(a) ResNet50 [11] with relatively higher Robustness Score on our Dataset, also perform good in MVTec AD dataset [2], can detect more details in scratches on metal nut rather than general focus on flaw shape

(b) VGG-16 [32] with relatively lower Robustness Score on our Dataset, perform poor in MVTec AD dataset [2]

Figure 4. Saliency Map Analysis on the MVTec AD Dataset [2].

Table 2. Variance of Model Accuracy Performance and Robustness Scores.

Model Acc. Drop Volatility	Scenarios								Variance	Robustness Score (Our* 0 - 1)	Offical Top-1 Acc. (On ImageNet [7])	Offical Top-1 Acc. (On Cifar10 [18])
	Blur_background	Blur_object	Image_g	Image_b	Image_grey	Image_r	Random_background	Segmented_image				
ResNet50 [11]: external	0,18%	0,50%	1,17%	0,68%	0,10%	0,38%	53,04%	7,12%	0,0902	0,8985	74,90%	93,03%
ResNet50 [11]: internal	0,10%	0,00%	0,11%	0,17%	0,20%	0,00%	6,96%	0,30%	0,0112			
DenseNet121 [13]:external	0,13%	0,72%	1,00%	1,01%	0,17%	0,84%	50,39%	7,69%	0,0885	0,9062	75,00%	96,54%
DenseNet121 [13]:internal	0,21%	0,00%	0,07%	0,14%	0,18%	0,01%	2,77%	0,29%	0,0052			
VGG-16 [32]:external	0,15%	0,15%	2,30%	5,45%	1,52%	1,72%	47,93%	19,53%	0,1125	0,8845	71,30%	93,43%
VGG-16 [32]:internal	0,33%	0,06%	0,26%	0,51%	0,72%	0,01%	0,01%	0,17%	0,0029			
ViT [8]:external	0,25%	0,07%	7,60%	9,38%	5,18%	7,24%	58,11%	19,74%	0,1536	0,8196	81,07%	98,20%
ViT [8]:internal	0,51%	0,72%	0,15%	0,07%	0,08%	0,57%	16,54%	0,00%	0,0266			
Swin [22]:external	0,96%	0,84%	6,84%	6,17%	4,61%	6,94%	50,87%	21,33%	0,1408	0,6305	83.58%	97,95%
Swin [22]:internal	0,02%	56,28%	0,48%	0,61%	0,56%	0,05%	37,10%	65,03%	0,2287			

Table 3. Performance Evaluation of Multiple Linear Regression. (P value < 0.0001 and **** indicate the result is of high significance. ns note as not significant).

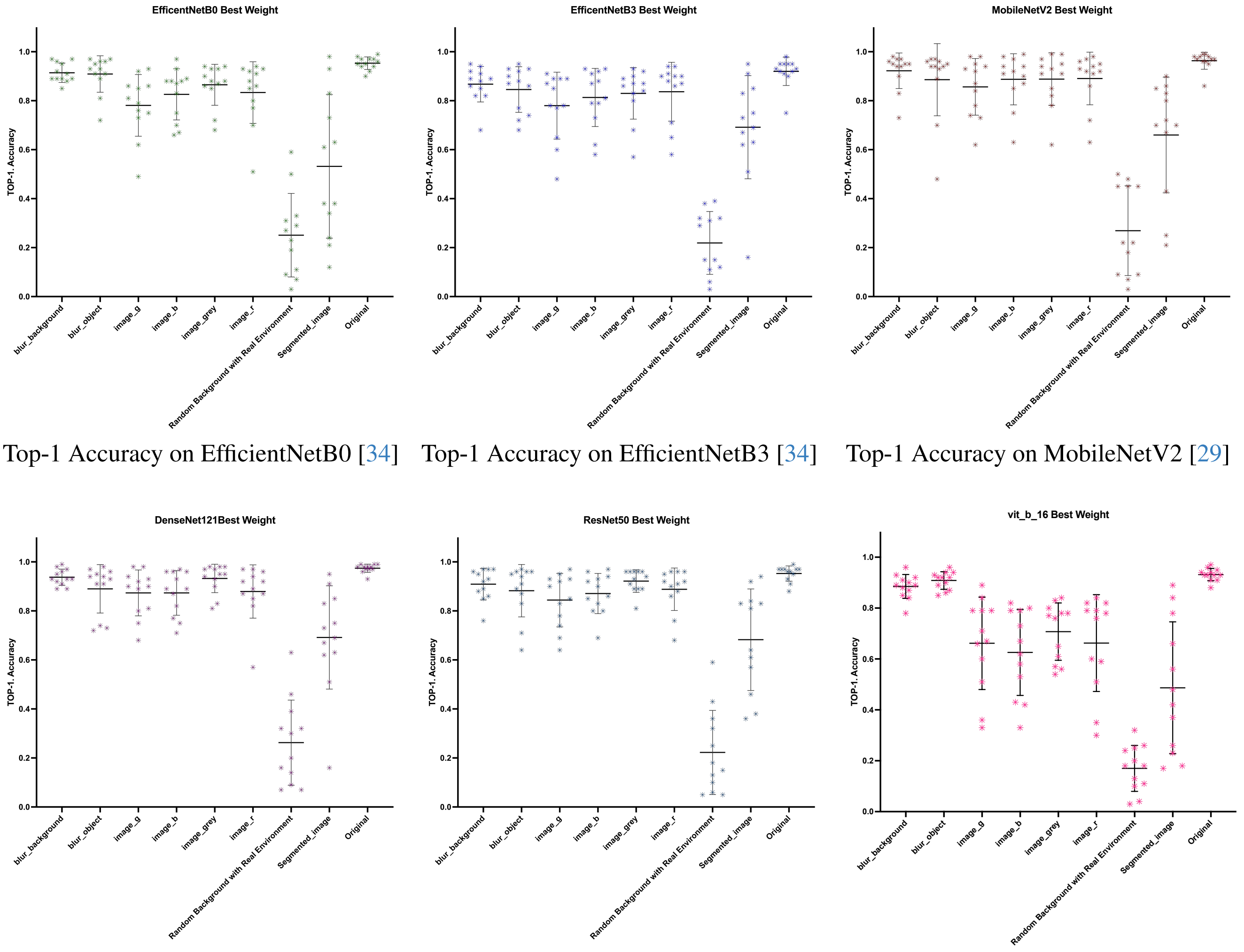
Classification				**Segmentation**			
Variable	**Estimate**	**P value**	**P value summary**	**Variable**	**Estimate**	**P value**	**P value summary**
Intercept	0.8986	< 0.0001	****	Intercept	**0.6672**	< 0.0001	****
Model Name[EfficientNetB0 [34]]	-0.03444	0.0175	*	Model Name[dpt_vit-b16 [8]]	**-0.1999**	< 0.0001	****
Model Name[EfficientNetB3 [34]]	-0.04111	0.0046	**	Model Name[upernet_swin [38]]	-0.2211	< 0.0001	****
Model Name[DenseNet121 [13]]	0.01556	0.2821	ns	Model Name[upernet_vit-b16 _ln_mln [38]]	-0.1695	< 0.0001	****
Model Name[MobileNetV2 [29]]	0.005556	0.7007	ns	Model Name[pspnet_r50-d8 [46]]	-0.045	0.0106	*
				Model Name[fpn_r50 [21]]	-0.2081	< 0.0001	****
				Model Name[upernet_r50 [38]]	-0.05796	0.001	**
Image Scenario[blur_background]	-0.0425	0.0288	*	Image Scenario[blur_background]	0.01833	0.3576	ns
Image Scenario[blur_object]	-0.07	0.0003	***	Image Scenario[blur_object]	-0.1571	< 0.0001	****
Image Scenario[image_g]	-0.1257	< 0.0001	****	Image Scenario[image_g]	-0.07131	0.0004	***
Image Scenario[image_b]	-0.0985	< 0.0001	****	Image Scenario[image_b]	-0.03952	0.0476	*
Image Scenario[image_grey]	**-0.06517**	0.0008	***	Image Scenario[image_grey]	-0.01929	0.3332	ns
Image Scenario[image_r]	-0.087	< 0.0001	****	Image Scenario[image_r]	-0.07702	0.0001	***
Image Scenario[Random Background with Real Environment]	**-0.7078**	< 0.0001	****	Image Scenario[segmented_image]	-0.08143	< 0.0001	****
Image Scenario[Segmented_image]	-0.3012	< 0.0001	****	Image Scenario[generated_background]	-0.1408	< 0.0001	****
Image Class[1]	0.134	< 0.0001	****	Image Class[1]	0.07619	0.001	***
Image Class[2]	-0.04867	0.0301	*	Image Class[2]	-0.06508	0.0048	**
Image Class[3]	0.04	0.0745	ns	Image Class[3]	0.05222	0.0234	*
Image Class[4]	0.1004	< 0.0001	****	Image Class[4]	0.08127	0.0004	***
Image Class[5]	0.1333	< 0.0001	****	Image Class[5]	0.2713	< 0.0001	****
Image Class[6]	0.07667	0.0007	***	Image Class[6]	0.3021	< 0.0001	****
Image Class[7]	0.01044	0.6409	ns	Image Class[7]	0.1641	7.137	****
Image Class[8]	0.09067	< 0.0001	****	Image Class[8]	0.1548	6.73	****
Image Class[9]	0.09933	< 0.0001	****	Image Class[9]	0.2216	9.635	****
Image Class[10]	0.1651	< 0.0001	****	Image Class[10]	0.2689	11.69	****
Image Class[11]	-0.02244	0.3164	ns	Image Class[11]	0.04079	1.774	ns

performance variance). Consistent with the observation in the previous part, ViT [8] and Swin-Transformer [22] have higher variances across diverse scenarios and lower robustness scores.

Table 2 also shows the performance on another dataset Cifar10 [18]. In this case, for CNN-based models, a model with a higher robustness score tends to have higher accuracy. For the transformer-based models, Swin-Transformer [22] has both a higher robustness score and better performance. This indicates that the robustness score calculated on XIMAGENET-12 can be an effective performance indicator for other datasets. To show that our robustness evaluation can also provide helpful guidance to real-world applications such as industry, we investigate the performance of ResNet50 [11] and VGG-16 [32] backbones on an industrial anomaly detection dataset MVTec AD [2]. We show the Saliency maps of ResNet50 [11] and VGG-16[32] on MVTec AD [2] in Figure 4. It indicates that the model with a higher robustness backbone tends to focus more on features from the foreground object and less on the background, such as scratches on metal nuts.

Performance Evaluation of Multiple Linear Regression. Here, we further examine the changes in model performance across scenarios (EX1), using multiple linear regression analyses. In addition to addressing classification tasks, we extend our investigation to include segmentation tasks.

As shown in Table 3 of multiple linear regression analyses, we consider these three variables: Model Name, Scenarios, and Image Class. We use the P-value to indicate the confidence of our results and use the P-value summary as an auxiliary indicator of the P-value.



Top-1 Accuracy on EfficientNetB0 [34] Top-1 Accuracy on EfficientNetB3 [34] Top-1 Accuracy on MobileNetV2 [29]

Top-1 Accuracy on DenseNet121 [13] Top-1 Accuracy on ResNet50 [11] Top-1 Accuracy on ViT [8]

Figure 5. TOP-1 Accuracy for SOTA models pre-trained on IMAGENET original formal images and tested on XIMAGENET-12 different background scenarios.

Specifically, the coefficient of Estimate for Segmentation Model Name[dpt-vitb16] is -0.1999, indicating that, compared to the reference segmentation Intercept model (deeplabv3plus-r50-d8 [5]), model[dpt-vitb16] is associated with a decrease of 0.1999 in segmentation accuracy. With most of the SOTA Visual segmentation model Accuracy decrease from -0.05 to -0.22 and the base Intercept model only achieve 0.6672 Accuracy with ($P < 0.0001$), it further verified that our benchmark could serve a valuable tool for measuring (SOTA) segmentation models performance in segmenting complex shapes or detecting detailed area in AI-generated background images (with Image Scenarios AI generated background leads to accuracy decrease by 0.14).

Besides, we can see that models suffer a performance drop once the background changes as all Image Scenarios Accuracy dropped from -0.7078 (Random Background) to color change (image_grey) -0.06517. Notably, the Classification(EX1) results in Table 3 also indicate ***foreground class also play an important role for content reasoning***, since Image Class[1,4,5] will lead to Accuracy increase in replacement of baseline Intercept Image Class[0], while other leads to decrease.

The results of the regression analyses are presented in Table 3, confirming our hypotheses. ***Our benchmark should present also a challenging task for SOTA segmentation models***. It serves as an effective tool for assessing model performance in segmenting complex shapes and detecting detailed areas within AI-generated background images.

Accuracy Drop of SOTA Models. We show the accuracy variance of SOTA visual models in Figure 5. Notably, we observed that the presence of "Random Background" had the most substantial adverse impact on accuracy, resulting in a significant decrease. Furthermore, the "Segmented Image" scenario also exhibited a significant negative influence, leading to a decrease in accuracy.

5.3. Qualitative Results

We compare the semantic annotations of our XIMAGENET-12 with the IOG benchmark dataset [48] and ImageNet-9 dataset [37] in Figure 6. As can be seen, the semantic labels of XIMAGENET-12 are much more precise than the others. We consider that due to the more precise separation of foreground and background, we can conduct a more reliable evaluation and analysis of the model robustness. For example, the sup-optimal annotation of ImageNet-9 [37] leads to a misleading claim that removing the background negatively impacts test accuracy. In contrast, we argue that poor segmentation quality, particularly with minimal foreground remaining, hampers the performance of recognition. We believe that our dataset can perform as a high-quality dataset for analysis of domain adaptation/generation.

We show the segmentation and attention map of SOTA segmentation models on XIMAGNET-12 in Figure 7. As can be seen, those segmentation models do not show satisfying performance. For example, the clothes of the dog has not been recognized. This indicates that our XIMAGNET-12 is a challenging dataset for the segmentation task.

6. DISCUSSION AND CONCLUSION

In this work, we introduce an explainable visual benchmark dataset, XIMAGENET-12, to evaluate the robustness of visual models. XIMAGENET-12 consists of six diverse scenarios, such as overexposure, blurring, color changes, etc., to simulate real-world situations. We further develop a robustness score to investigate the model performance under various conditions. From the experiments, we conclude the following interesting findings:

1) Different scenarios influence visual models in different degrees, and randomly substituting the background leads to the most severe performance drops.

2) Models trained and tested with well-segmented foregrounds tend to perform well even if the backgrounds are missing.

3) A model with higher accuracy is not necessarily more stable.

We expect the XIMAGENET-12 dataset will empower researchers to thoroughly evaluate the robustness of their visual models under challenging conditions. In future work, we will show how XIMAGENET-12 can serve as a high-quality dataset for more visual applications such as semantic segmentation tasks and domain adaptation/generation tasks.

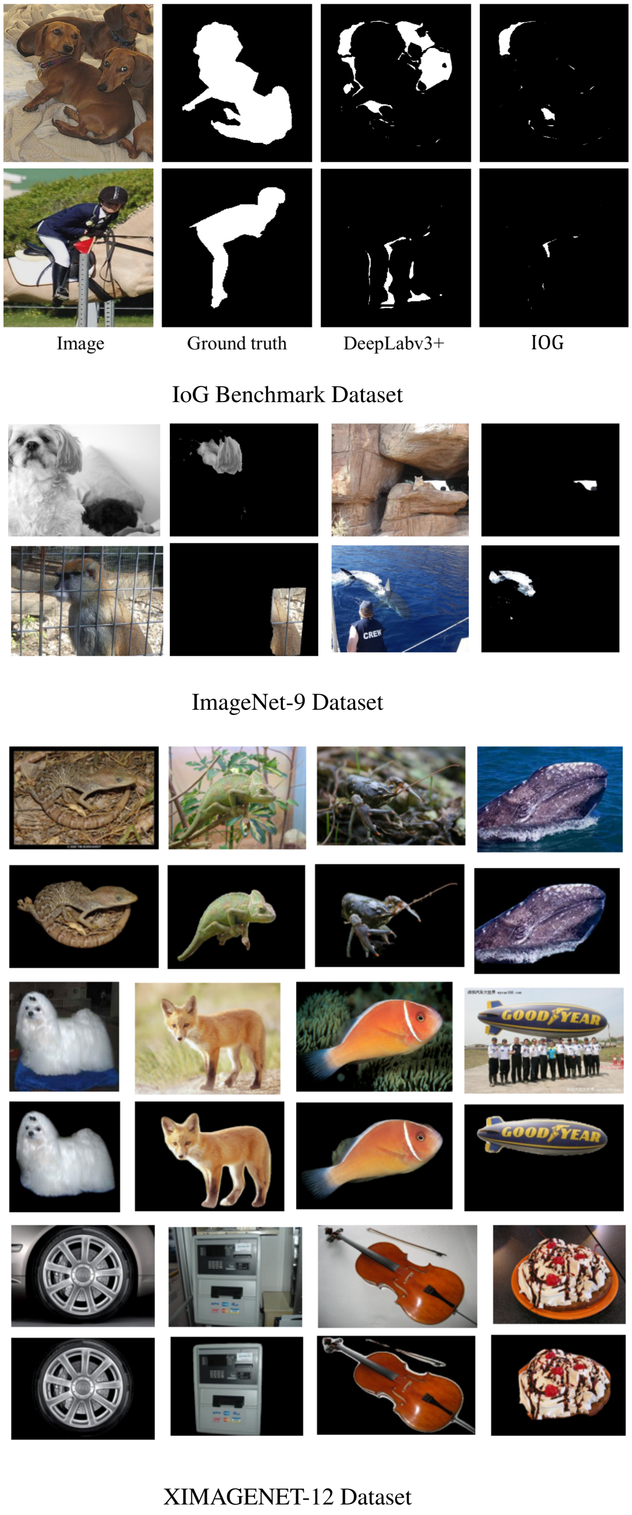


Figure 6. Additional Related Works and Explicit Comparisons. As can be seen, the semantic annotation of XIMAGENET-12 dataset is much more precise than the other datasets.

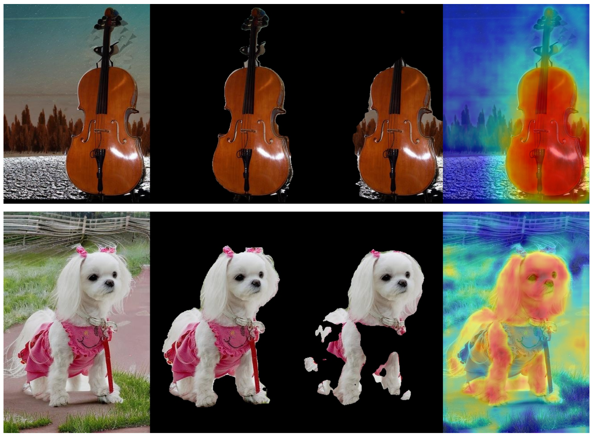


Figure 7. Attention Map of SOTA segmentation models. "pspnet [46] r50-d8" achieves a MIoU of 0.562 in our dataset. Our dataset is also enabled to tackle the intricacies of segmenting objects within AI-generated backgrounds, offering a substantial improvement in human-labeled ground truth quality compared to the original ImageNet [7], where only image labels are included. Moreover, it proves to be a valuable resource for identifying AI-generated images on the internet, showcasing its versatility and significance in contemporary computer vision research.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. 4

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, pages 9592–9600, 2019. 5, 6

[3] Hanning Chen, Wenjun Huang, Yang Ni, Sanggeon Yun, Fei Wen, Hugo Latapie, and Mohsen Imani. Taskclip: Extend large vision-language model for task oriented object detection, 2024. 1

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 4

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018. 4, 7

[6] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 4

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 2, 3, 5, 6, 9

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4, 5, 6, 7

[9] Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer, 1970. 4

[10] Haonan Han, Rui Yang, Shuyan Li, Runze Hu, and Xiu Li. Ssgd: A smartphone screen glass dataset for defect detection. In *ICASSP*, pages 1–5, 2023. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4, 5, 6, 7

[12] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *CVPR*, pages 1314–1324, 2019. 5

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 4, 5, 6, 7

[14] Ziyu Jia, Youfang Lin, Yuhan Zhou, Xiyang Cai, Peng Zheng, Qiang Li, and Jing Wang. Exploiting interactivity and heterogeneity for sleep stage classification via heterogeneous graph neural network. In *ICASSP*, 2023. 1

[15] Nikhil Ketkar and Eder Santana. *Deep learning with Python*. Springer, 2017. 4

[16] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A. Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *ECCV*, pages 158–171, 2012. 2

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 25, 2012. 5

[20] Qiang Li and Chongyu Zhang. Continual learning on deployment pipelines for machine learning systems. In *NeurIPS*, 2022. 1

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 4, 6

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *CVPR*, pages 10012–10022, 2021. 4, 5, 6

[23] Evrim Ozmermer and Qiang Li. Self-supervised learning with temporary exact solutions: Linear projection. In *INDIN*, pages 1–7. IEEE, 2023. 1

[24] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *NeurIPS*, 2017. 4

[25] PlaygroundAI. playgroundai, 2023. 3

[26] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3

[27] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *SIGKDD*, pages 1135–1144, 2016. 1

[28] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *CoRR*, abs/1911.08731, 2019. 1

[29] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018. 4, 5, 6, 7

[30] Andrew D Selbst and Solon Barocas. The intuitive appeal of explainable machines. *Fordham L. Rev.*, 87:1085, 2018. 1

[31] Rakshith Shetty, Bernt Schiele, and Mario Fritz. Not using the car to see the sidewalk - quantifying and controlling the effects of context in classification and segmentation. In *CVPR*, pages 8218–8226. Computer Vision Foundation / IEEE, 2019. 2

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5, 6

[33] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015. 4

[34] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114. PMLR, 2019. 4, 5, 6, 7

[35] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, pages 1521–1528. IEEE, 2011. 2

[36] V7Drawin. V7drawin, 2023. 3

[37] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *CVPR*, 2020. 1, 2, 5, 8

[38] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, pages 418–434, 2018. 4, 6

[39] G Udny Yule. On the theory of correlation. *Journal of the Royal Statistical Society*, 60(4):812–854, 1897. 4

[40] Oliver Zendel, Wolfgang Herzner, and Markus Murschitz. Vitro-model based vision testing for robustness. In *IEEE ISR*. IEEE, 2013. 5

[41] Dan Zhang, Fangfang Zhou, Yuwen Jiang, and Zhengming Fu. Mm-bsn: Self-supervised image denoising for real-world with multi-mask based on blind-spot network. In *CVPR*, pages 4188–4197, 2023. 1

[42] Dan Zhang, Fangfang Zhou, Felix Albu, Yuanzhou Wei, Xiao Yang, Yuan Gu, and Qiang Li. Unleashing the power of self-supervised image denoising: A comprehensive review, 2024. 1

[43] Jianguo Zhang, Marcin Marszalek, Svetlana Lazebnik, and Cordelia Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 73(2):213–238, 2007. 1

[44] Jie Zhang, Masanori Suganuma, and Takayuki Okatani. Network pruning and fine-tuning for few-shot industrial image anomaly detection. In *INDIN*, pages 1–6, 2023. 1

[45] Shiyin Zhang, Jun Hao Liew, Yunchao Wei, Shikui Wei, and Yao Zhao. Interactive object segmentation with inside-outside guidance. In *CVPR*, pages 12234–12244, 2020. 3

[46] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 4, 6, 9

[47] Fangfang Zhou, Zhengming Fu, and Dan Zhang. High dynamic range imaging with context-aware transformer. In *IJCNN*, pages 1–8. IEEE, 2023. 1

[48] Zhuotun Zhu, Lingxi Xie, and Alan L. Yuille. Object recognition with and without objects. In *IJCAI*, pages 3609–3615. ijcai.org, 2017. 1, 2, 8