# Robustness Analysis on Foundational Segmentation Models

Madeline Chantry Schiappa[1] Shehreen Azad [1*] Sachidanand VS[2]
Yunhao Ge[3] Ondrej Miksik[4] Yogesh S Rawat[1] Vibhav Vineet[4]

[1]Center for Research in Computer Vision, University of Central Florida
[2]Indian Institute of Technology, Madras; [3]University of Southern California; [4]Microsoft Research

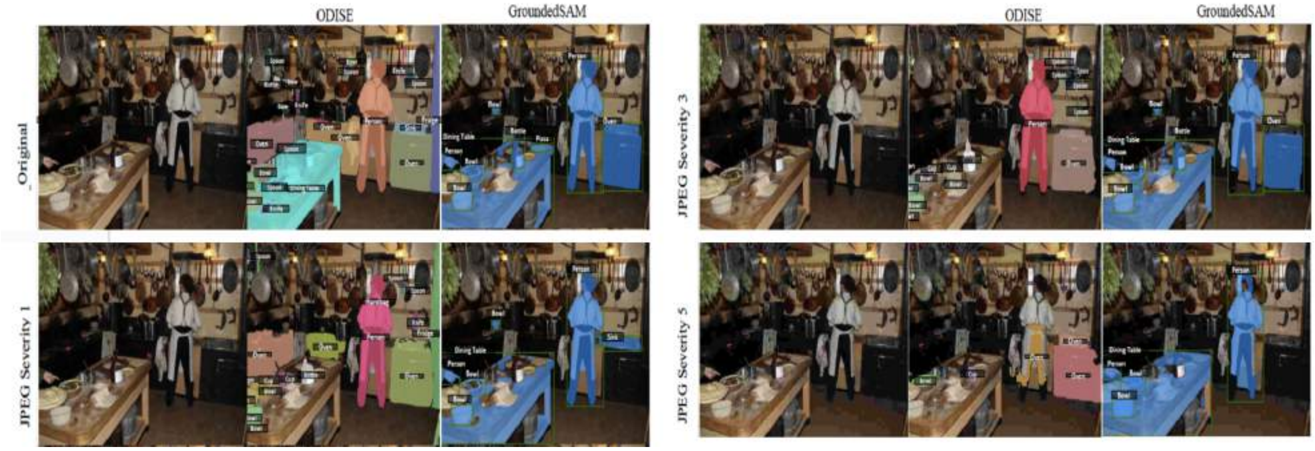

Figure 1. **Segmentation with foundation models** ODISE [56] and GroundedSAM [45] on corrupted images (JPEG compression) at varying severity. An interesting observation is with corruption, while the person is still clearly visible, ODISE fails to recognize it even at severity 1.

## Abstract

*Due to the increase in computational resources and accessibility of data, an increase in large, deep learning models trained on copious amounts of multi-modal data using self-supervised or semi-supervised learning have emerged. These "foundation" models are often adapted to a variety of downstream tasks like classification, object detection, and segmentation with little-to-no training on the target dataset. In this work, we perform a robustness analysis of Visual Foundation Models (VFMs) for segmentation tasks and focus on robustness against real-world distribution shift inspired perturbations. We benchmark seven state-of-the-art segmentation architectures using 2 different perturbed datasets, MS COCO-P and ADE20K-P, with 17 different perturbations with 5 severity levels each. Our findings reveal several key insights: (1) VFMs exhibit vulnerabilities to compression-induced corruptions, (2) despite not outpacing all of unimodal models in robustness, multimodal models show competitive resilience in zero-shot scenarios, and (3) VFMs demonstrate enhanced robustness for certain object categories. These observations suggest that our robustness evaluation framework sets new requirements for foundational models, encouraging further advancements to bolster their adaptability and performance. The code and dataset is available at:* `https://tinyurl.com/fm-robust`.

## 1. Introduction

Visual Segmentation is a longstanding challenge in computer vision, encompassing various tasks. These tasks require varying degrees of detail and include semantic [6, 36, 63], panoptic [28, 62], instance [20, 30, 53]. Traditionally, different tasks and datasets were handled independently with specialized models [20, 28, 36, 42], which did not allow cross-task synergy. However, with the advent of versatile transformer-based models [14, 52] and large-scale vision-language pre-training [4, 7, 21, 44], there's a growing shift towards developing comprehensive, multi-purpose, open-vocabulary vision systems, known as Visual Foundation Models (VFMs) [31, 34, 61].

In addition, inspired by the success of Large Language Models (LLMs), such as ChatGPT [44], VFMs have harnessed the immense potential of foundation models and

*Corresponding Author: Shehreen.Azad@ucf.edu adapted them to open vocabulary instance segmentation tasks. For example, VFMs like Segment Anything (SAM) [29], ODISE [56] possesses the ability to segment any object within images without the requirement for further training. Such a breakthrough has ushered in many new opportunities in many safety critical real-world applications including autonomous vehicles, healthcare systems, etc. [13, 17, 37, 58].

Deploying models in real world often introduces distribution shifts in the data, leading to unforeseen model behavior. To address this, studying the robustness of current deep learning models against potential real-world perturbations is essential [11, 16, 22, 26, 48, 50]. These perturbations are not artificially induced through adversarial attacks, but naturally occurred due to changes in the environment, varying camera settings, and compression. Hendrycks et.al. [22] introduced a series of such perturbations and evaluated the robustness of image classification models to these perturbations. Following this, such perturbations have been applied to evaluate robustness of models in several other downstream tasks [27, 49]. While they studied the robustness of models in supervised settings for classification tasks; the robustness of VMFs for segmentation tasks remains uncertain regardless of the type of supervision during learning. As VFMs become increasingly common and adapted for numerous downstream tasks, understanding their robustness and behavior in response to potential real-world distribution shifts is crucial.

In this work, we conduct an extensive robustness analysis of VFMs with billions of parameters for the segmentation tasks. We use four recent multimodal VFM-based models, namely ODISE [56], Painter [54], InternImage [53], Segment-Anything (SAM) [29] along with recent unimodal models, namely Mask2Former [10], MaskDINO [30] and ViT-Adapter [8]. For the first two non-VFM models, we use both CNN and transformer based backbones. For the robustness analysis, we use 17 common perturbations with 5 severity levels to the MS COCO [32] and ADE20K [64] datasets and name the perturbed datasets as MS COCO-P and ADE20K-P for the task of segmentation.

Our findings indicate that from the studied models, (1) VFMs lack robustness in compression and blur based corruptions. (2) All of the multimodal VFMs are not noticeably more robust nor higher performing in these segmentation tasks than the unimodal models; but have competitive robustness in a zero-shot setting. (3) multimodal VFMs show higher relative robustness for specific object-types compared to unimodal modelas.

In summary, our contributions are as follows:

- We focus on robustness analysis of foundational segmentation models against distribution shifts due to real-world inspired perturbations.
- We provide two benchmark datasets (MS COCO-P and ADE20K-P) to conduct robustness analysis on segmentation tasks.
- We present an empirical analysis of foundational modeling approaches in segmentation to study the effect of various perturbations on their performance.

## 2. Related Work

### 2.1. Vision Foundation Models

The field of AI has seen a paradigm shift with the emergence of models trained on massive amounts of data at scale and are adaptable to various downstream tasks; which are commonly referred to as foundation models [1, 40, 43, 47]. These models have shown remarkable performance in language and vision-related tasks, e.g. retrieval, recognition, segmentation etc. Recent works on multi-modal learning have a trend of embedding features from different type of inputs to a common feature space [1, 43]; which is achieved by training these models using contrastive learning. Due to this common feature space of text in image embedding, it has been used for a huge number of downstream tasks [41, 47, 60]. Stable diffusion [47] is one such popular model commonly used for generative purposes to solve the downstream tasks. To overcome the drawback of CLIP which overlooks the visual local information, DINOv2 [40] is proposed, which is trained with self-supervised learning. A closed-set detector Dino [5] is extended into Grounding DINO [33] for open-set object detection by performing vision-language modality fusion at multiple phases.

Segmentation is an important computer vision task with safety-critical applications such as medical imaging and self-driving scenarios where robust models are required since wrong results could be catastrophic depending on the situation. There are some foundation models which have been developed for specific segmentation related tasks [29, 53, 54, 65] some of which are based on the aforementioned models. However, while deployed in the real world the data these models come across might be corrupted with different kind of distribution shift, thus creating a necessity for robustness benchmarking for these models.

### 2.2. Robustness

Recently many work has focused on evaluating the vision model's robustness in image [2, 22, 24, 26, 51] and video domain [49, 50]. Hendrycks et al. [22] showed that the performance of bigger models gets affected just as the smaller models by corrupted data on the task of image classification. This shows that even though larger models may have more capacity to capture intricate features, they are not immune to the challenges posed by dataset variations or perturbations, leading to similar performance impacts as observed in smaller models. Following these works, such data corruptions were used for evaluating robustness of models in classification and object detection tasks. [38, 51]. The robustness of segmentation models has also been explored in recent works [26] using similarly perturbed dataset for evaluating the robustness of segmentation models highlighting a significant performance drop for corruptions affecting image texture versus those preserving it. There are similar work on video domain [50] also evaluating the robustness of video action recognition models. There has also been work on improving model's robustness using augmentation techniques [18, 24, 38, 59].

With the growing popularity of the foundation model, performing very well in many areas of computer vision and with many downstream tasks being solved by using foundation model as an encoder, it is important to understand their behavior and robustness to potential real-world distribution shifts in the data. Towards this goal, we use a set of perturbations that are frequently encountered in real-world environments on datasets designed for segmentation tasks and evaluate the robustness of multi-modal models and compared it with unimodal models.

# 3. Experiments and Results

## 3.1. Distribution Shifts and their Severity

Corruptions due to adversarial attacks are intentionally crafted to exploit vulnerabilities in machine learning models by adding imperceptible perturbations to the input data, thus causing misclassification or incorrect predictions. Unlike data corruption due to adversarial attack real-world data corruptions affect data during capture, transmission, or storage, and can degrade its quality.

In this work we study six different categories of real-world perturbations typically used in robustness benchmarking [22, 23, 27, 48, 49]. These categories include noise, blur, compression, digital, camera, and environmental perturbation and there is a total of *17* different perturbations across all these categories. In noise, we have *gaussian*, *shot*, *impulse*, and *speckle* noise. In the blur category, we have *defocus*, *motion*, and *zoom* blur. In the compression category, we have *jpeg* and *pixelate* corruption. In the digital category, we have *contrast* and *shear*. In the camera category, we have *translate* and *rotate* and finally, in the environment category, we have *brightness*, *darkness*, *snow*, and *fog*. The algorithms used to generate these corruptions follow previous literature [22, 27, 48].

In the real world, distribution shift corruptions may occur in varying levels of severity depending on the environment and/or situation. Therefore it is important to evaluate models under the same assumption that corruptions can vary in severity. We generate five levels of severity where 1 is a small shift and 5 is a large distribution shift (Figure 2). We apply all the proposed corruptions for each severity on all images using the *imgaug* [25] library and code available from [48] to generate the corruptions and their corresponding annotation.

## 3.2. Model Variants

We perform our benchmark evaluation on seven state-of-the-art methods. We selected a set of models that were representative of multimodal Visual Foundation Models (VFMs) (ODISE [56], two variations of Segment-Anything [29]- PromptSAM [46] and GroundedSAM [45], Intern-Image [53], Painter [54]) and comparative state-of-the-art unimodal models (ViT-Adapter [8], Mask2Former [10] and MaskDINO [30])for segmentation. We selected models based on their availability of code, weights, and reproducability.

**ODISE** [56] is based on the feature space learned in Stable Diffusion [47] for their image encoder, CLIP [43] for their image-text discriminator and Mask2Former [10] as their mask generator. The process starts with extracting image features from Stable Diffusion with a *Implicit Captioner* to learn implicit text prompts. These embeddings are passed to the mask generator. Similarity is measured between each mask and text embeddings of object categories from CLIP [43] to assign a class to a mask. While the model is trained on one dataset, it can be applied to any dataset for zero-shot evaluation, making it a strong model to consider.

**Segment-Anything** (SAM) [29] uses a MAE [21] pre-trained Vision Transformer (ViT) [15] image-encoder and a set of prompts that are either points, text, or bounding boxes to mask desired objects. This model is also designed for zero-shot transfer. We adopt two variants : **PromptSAM** [46] and **GroundedSAM** [45]. PromptSAM uses FocalDINO [5, 57] to generate bounding box proposals as prompts. GroundedSAM uses GroundingDINO [33] to generate open-vocabulary bounding boxes as prompts to SAM. Because GroundingDINO cannot generate discriminate prompts for "stuff" categories for semantic segmentation tasks, we only evaluate on instance segmentation.

**InternImage** [53] is a large-scale CNN-based foundation model which uses deformable convolution as its core operator. It reduces the strict inductive bias of traditional CNNs and makes it possible to learn stronger and more robust patterns with large-scale parameters from massive data like ViTs. For our experiments, we use InternImage-XL with cascade method for MS COCO-P and InternImage-H with UperNet framework for ADE20K-P evaluation due to availability of code and weights.

**Painter** [54] is a generalist model which implements in-context learning [3] in NLP to vision tasks. They redefine the output format of chosen tasks to image format and use a masked autoencoder based approach for training the model. Painter has achieved competitive performance compared to well-established task-specific models, which makes it a very powerful model for our task.

We compare these 5 multimodal VFM based approaches to

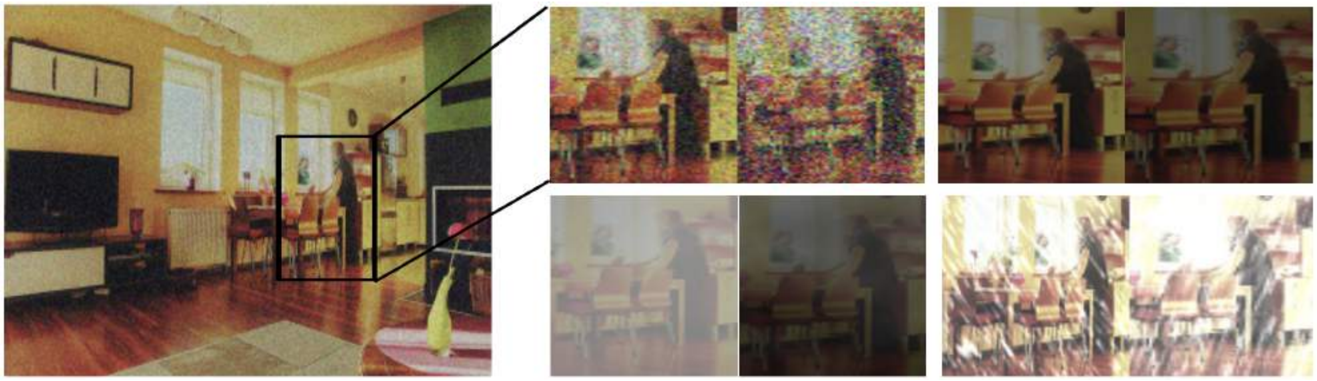

Figure 2. **Data perturbation examples** where original sample is zoomed in to show different corruptions on image from the MS COCO-P dataset. Each image pair is of corruption at severity 3 and 5. Top row shows corruptions in the category of gaussian noise and darkness, whereas, bottom row shows fog and snow.

Table 1. **Architectural details of the models** used for robustness analysis and their size and the number of parameters used for fine-tuning (FT) the model on the evaluation datasets. Here, Param-M: model parameters in millions, T-Param-M: trainable model parameters in millions, FT-C: finetuned on COCO, FT-A: finetuned on ADE20K

| | Model | Backbone(s) | Param-M | T-Param-M | FT-C | FT-A |
|---|---|---|---|---|---|---|
| Unimodal | MaskDINO [30] | R50[19] | 53.25 | 53.25 | True | True |
| | MaskDINO [30] | SwinL[35] | 224.39 | 224.39 | True | True |
| | Mask2Former [10] | R50[19] | 44.00 | 44.00 | True | True |
| | Mask2Former [10] | SwinL[35] | 216 | 216 | True | True |
| | Vit-Adpater-L [8] | ViT[15] | 348 | 348 | True | False |
| Multimodal | InternImage [53] | InternImage-XL[35] | 387 | 387 | True | False |
| | Painter [54] | ViT[15] | 371 | 371 | True | True |
| | PromptSAM [29, 46] | FocalDINO[5, 57],MAE+ViT | 321.86 | 228.12 | True | _ |
| | GroundedSAM [29, 33] | GroundingDINO[33],MAE[21]+ViT[15] | 834.99 | 232.90 | True | False |
| | ODISE [56] | Mask2Former[10], CLIP[43], GLIDE[39] | 1,521.90 | 28.10 | True | False |

3 unimodal based methods.

**MaskDINO** modifies DINO [5], a self-supervised approach using self-distillation. ResNet50 (R50) [19] and SwinL [35] backbones are used in our evaluation for MaskDINO.

**ViT-Adapter** is a pre-training free additional network that can efficiently adapt the plain ViT [15] to downstream dense prediction tasks without modifying its original architecture.

**Mask2Former** is a modified MaskFormer [9], which uses a transformer-based module that produces per-segment embeddings and a pixel-decoder module that produces per-pixel embeddings. The pixel-decoder module uses a backbone of either ResNet50 or SwinL. Table 1 presents more details about all of our used models.

### 3.3. Datasets

We use two segmentation benchmark datasets for our experiments: MS COCO Panoptic [32] and ADE20K [64]. MS COCO dataset has 80 "things" categories and 53 "stuff" categories. ADE20K has 100 "things" and 50 "stuff" categories. For each dataset, we perturb an image with each of the 17 corruptions and 5 severities, resulting in 425,000 images for the **MS COCO-P** dataset and 170,000 images for the **ADE20K-P**.

### 3.4. Benchmark Evaluation Metrics

**Performance Metrics:** We evaluate the models on instance and semantic segmentation tasks for our MS COCO-P and ADE20K-P dataset. Each datasets has a category of ”things” and ”stuff” categories; in which ”things” are countable objects like people, animals, etc; whereas ”stuff” are amorphous regions like sky, grass, etc. *Semantic segmentation* is evaluated on both the "things" and "stuff" categories using mean intersection over union (mIoU). *Instance segmentation* is only evaluated on the "things"category using mean average precision (mAP) on the "things" categories. For models trained on panoptic segmentation, all masks assigned to one ”thing” category are merged into a single mask.

**Robustness Metrics:** To measure robustness we use two metrics: absolute and relative robustness [50]. We start by measuring the performance of a trained model $f$ on a clean set of data $A_c^f$ and a corrupted data $A_{p,s}^f$. Here, $A_{p,s}^f$ is corrupted by perturbation $p$ at each severity level $s$. Relative robustness ($\gamma_{p,s}^r$) measures the relative drop in performance between original samples and a corrupted sample, whereas, absolute robustness $\gamma_{p,s}^a$ measures the absolute drop in performance. These can be computed as eqn. 1 and eqn. 2.

$$\gamma_{p,s}^r = 1 - \frac{A_c^f - A_{p,s}^f}{A_c^f} \tag{1}$$

$$\gamma_{p,s}^a = 1 - \frac{A_c^f - A_{p,s}^f}{100} \tag{2}$$

Because $\gamma^a$ will also depend on the models performance on clean images, we emphasize $\gamma^r$, focusing on the change in performance. Nevertheless, the results of $\gamma^a$ is reported in the supplementary. These metrics are calculated between $0-1$, where higher score means more robustness.

**Implementation Details:** All models are used in accordance to the provided code and model weights. The models Mask2Former, ODISE and MaskDINO all rely on Detectron2 [55] evaluation code. For SAM evaluation, the package *mmsegmentation* [12] was used . For GroundedSAM, we were unable to replicate results for its bounding box detector GroundingDINO. While we do provide results for this model, please note that we did our best to replicate given there was no evaluation code provided for either datasets. On the ADE20K dataset, all ODISE and SAM-based models are evaluated zero-shot whereas other models are trained.

### 3.5. Results

The relative robustness $\gamma^r$ and absolute robustness $\gamma^a$ scores for *instance segmentation* and *semantic segmentation* on MS COCO-P and ADE20K-P is shown respectively in Table 2 and Table 3, where each row corresponds to the average robustness across all corruptions and severity. Here, robustness on each category of segmentation is reported for only those models that had publicly available weights and code for the selected dataset for the selected task. Additional results across all perturbation category for both $\gamma^r$ and $\gamma^a$ is reported in the supplementary.

## 4. Analysis and Discussion

**All models struggle with blur and compression:** The relative robustness $\gamma^r$ scores and mean average precision (mAP) scores for *instance segmentation* for MS COCO-P for all distribution shifts is shown respectively in Figure 3 and Figure 5. We observe that the selected models are typically robust to all shifts with the exception to blur and compression. Figure 1 shows an example of compression corruptions for ODISE and GroundedSAM. Here, even for compression at severity level 5, the objects are clearly visible for the human eye. However, ODISE is struggling to properly classify objects even at severity 1 (person is denoted as handbag too), even though the generated mask is correct. However, as the severity increases, even though the object boundary is clear to the naked human eye, ODISE's generated mask gets more degraded in quality. GroundedSAM on the other hand generates correct masks and accurate recognition even at severity 5.

Table 2. **Absolute ($\gamma^a$) and Relative ($\gamma^r$) robustness scores for MS COCO-P**, where higher values mean more robust, averaged across all corruptions and severity. Here, *IS* and *SS* denotes instance and semantic segmentation respectively.

| | IS | | SS | |
|---|---|---|---|---|
| | $\gamma^a$ | $\gamma^r$ | $\gamma^a$ | $\gamma^r$ |
| Mask2Former+R50 | 0.86 | 0.68 | 0.85 | 0.75 |
| MaskDINO+R50 | 0.86 | 0.68 | 0.85 | 0.74 |
| Mask2Former+SwinL | 0.91 | 0.81 | 0.94 | 0.92 |
| MaskDINO+SwinL | 0.91 | 0.81 | **0.95** | **0.92** |
| ViT-adapter-L | 0.91 | 0.80 | – | – |
| ODISE+Label | 0.90 | 0.79 | 0.92 | 0.88 |
| ODISE+Caption | – | – | 0.93 | 0.87 |
| Prompt+SAM | 0.92 | 0.81 | – | – |
| InternImage-XL | 0.91 | 0.81 | – | – |
| PAINTER | **0.95** | **0.82** | 0.92 | 0.87 |
| GroundedSAM+SwinB | 0.92 | 0.80 | – | – |

Table 3. **Absolute ($\gamma^a$) and Relative ($\gamma^r$) robustness scores for ADE20K-P**, where higher values mean more robust, averaged across all corruptions and severity. Here, *IS* and *SS* denotes instance and semantic segmentation respectively.

| | IS | | SS | |
|---|---|---|---|---|
| | $\gamma^a$ | $\gamma^r$ | $\gamma^a$ | $\gamma^r$ |
| Mask2Former+R50 | 0.89 | 0.57 | 0.84 | 0.65 |
| MaskDINO+R50 | – | – | 0.82 | 0.63 |
| Mask2Former+SwinL | 0.94 | **0.92** | 0.93 | 0.87 |
| ViT-adapter-L | – | – | 0.94 | 0.89 |
| ODISE+Label | **0.97** | 0.79 | **0.97** | **0.89** |
| InternImage-H | – | – | 0.93 | 0.87 |
| PAINTER | – | – | 0.92 | 0.83 |
| GroundedSAM+SwinB | 0.95 | 0.73 | – | – |

In summary, while all models struggle with *blur and compression* corruptions, ODISE and SAM are the particularly lower performing ones in terms of robustness. Since both of these foundation models use a generative model for image embeddings, this could be a reflection of generative model robustness to compression based corruptions. Figure 5 presents the performance of all models in terms of mean Average Precision (mAP). Interestingly, the model with the highest robustness score, PAINTER, exhibits the lowest mAP. Despite its lower mAP compared to other models, PAINTER demonstrates superior robustness in terms of $\gamma^r$ score across various corruption categories. This high-

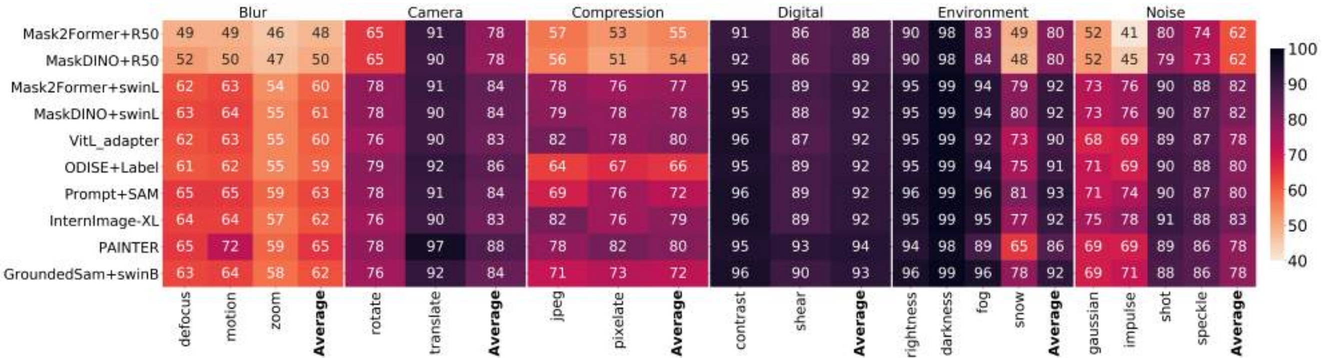

Figure 3. **Relative robustness score $\gamma^r$ on instance segmentation** for the MS COCO-P dataset. Here the Y-axis denotes the models we evaluated and the x-axis denotes $\gamma^r$ for each corruption averaged over severity.

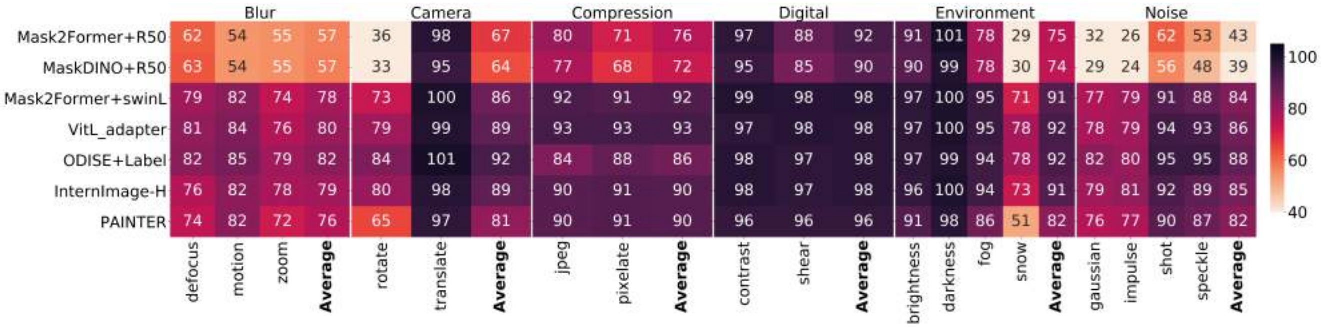

Figure 4. **Relative robustness score $\gamma^r$ on semantic segmentation on ADE20K-P**. Here, the Y-axis denotes the models we evaluated and x-axis denotes $\gamma^r$ for each corruption averaged over severity.

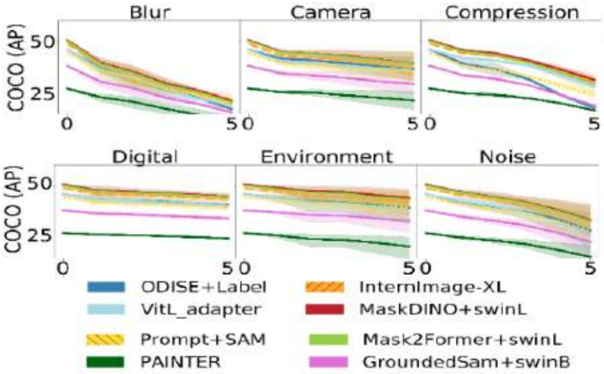

Figure 5. **mAP score on instance segmentation on MS COCO-P**. Here x axis denotes severity ranging from 0 (no corruption) to 5 (severe corruption) and y axis denotes the model performance measured by mean average precision (mAP).

lights a trade-off between selecting a model based on its performance versus its robustness. In this case, PAINTER's resilience to corruption shifts makes it a compelling choice despite its slightly lower overall performance in mAP.

**Multimodal models are not typically more robust or higher performing; but are consistent on zero-shot:** Since traditionally ADE20K is more popularly used for semantic segmentation than instance segmentation, we report the relative robustness $\gamma^r$ scores for ADE20K-P dataset across all categories for semantic segmentation in Figure 4. In Table 2, Table 3, Figure 3 and Figure 4, we show robustness results across model types defined by whether a model is a "foundation" model or "non-foundation" model, and whether the non-foundation model uses a CNN or a transformer based backbone. Now, these results do not provide convincing evidence that all multimodal models are typically more robust than unimodal. However, the most robust model is indeed a multimodal model, but this does not necessarily prove multimodal models' strength over unimodal models. There is a more noticeable difference for absolute robustness (Table 3) on the ADE20K-P datset where foundation models are evaluated zero-shot. So while models may not be typically more robust or higher performing, their zero-shot capability allows for much greater flexibility. When observing Figure 6, we see that even for hard cases of corruption like blur and compression, even as the severity increases, the zero shot performance of the multimodal models remain relatively stable in comparison to unimodals. This same observation is seen in case of robustness as well (Figure 4).This consistent performance of multimodal models is seen across all different categories of cor-

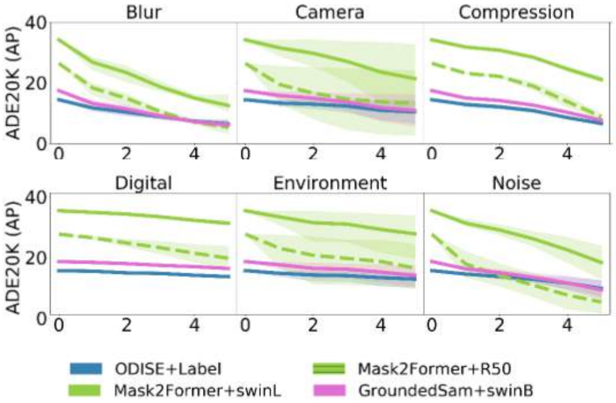

Figure 6. **mAP score on instance segmentation on ADE20K-P**. Here x axis denotes severity and y axis denotes the model performance measured by mean average precision (mAP). All ODISE and SAM-based models are evaluated zero-shot.

ruption. In summary, while the selected multimodal models are not typically more robust or higher-performing than the unimodal ones, they show promising zero-shot capabilities that have competitive robustness scores across both instance and semantic segmentation tasks.

**All models lack robustness in texture non-preserving corruptions:** Even though as per Figure 4, we observe that models show robustness across the corruption category as well for the ADE20K-P dataset; from Figure 3, we see almost all models performance are affected by all corruptions in the blur, compression category; snow from the environment category, gaussian, and impulse from the noise category of corruptions for the MS COCO-P dataset. Apart from compression, robustness drop for corruptions due to all the aforementioned categories are also valid in case of the ADE20K-P dataset. These results align with previous works [18, 26] that distortions that corrupt the texture of an image have a negative effect on model robustness compared to texture-preserving corruptions such as brightness, contrast, and geometric corruption. This shows that both multimodal and unimodal models are not robust to distortions that corrupts image texture. We have finetuned a few models on an augmented dataset consisting of these texture non-preserving corruption to see how it affects performance.

In this experiment, we fine-tune the models InternImage-XL, Vit-adapter-L, and Mask2Former+swinL using an augmented dataset specific to a particular category. The objective is to assess whether these augmentations enhance the models' robustness. We evaluate the models on the ADE20K-P dataset for semantic segmentation task, focusing on their performance under varied perturbations. We present the results of the InternImage model in Figure 7, whereas results of other models on the augmented dataset and more details about the fine-tuning dataset is provided in the supplementary. Across all models, augmentations generally elevate the perturbation score of their respective category, except for compression augmentation in Mask2Former, which adversely affects performance across all categories, including compression. Notably, blur and noise augmentation substantially elevate robustness scores in each model's relevant category.

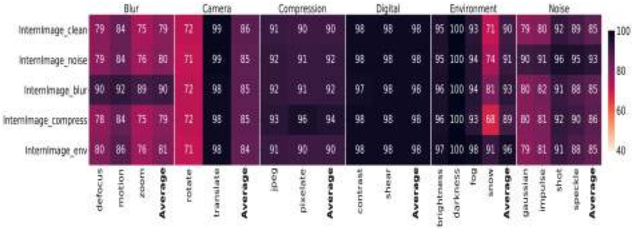

Figure 7. **Fine-tuned performance of InternImage on semantic segmentation for the ADE20K-P dataset**. Y-axis: Augmentation used for fine-tuning (expect first row). X-axis: model Relative Robustness score for each corruption averaged over severity.

**Multimodal models are relatively more robust to certain objects; especially under blur and compression:** To evaluate how model performance per object is impacted under different corruptions, we evaluate per-object relative robustness ($\gamma^r$) scores. Figure 8 shows a summary of the $\gamma^r$ across 11 super-categories for objects and corruptions under each distribution shift category for the MS COCO-P dataset. For mapping of object to super category, original MS COCO documentation is followed.

When looking at super-categories in Figure 8, we observe that multimodal models demonstrate greater relative robustness for certain object categories for certain perturbations. This is most noticeable under compression, blur and noise for objects in "outdoor" and "sports". To better understand these patterns, Table 4 reports the average relative robustness ($\gamma^r$) scores of each object super category across all corruptions and severity. From this table, we observe the selected multimodal models are typically more robust against objects under "appliance", "furniture", "outdoor" and "sports" when averaged across all corruptions and severity. While objects in furniture and outdoor tend to be quite large, the objects in sports are quite small. Therefore the size of the objects may not be the factor for this robustness. However, why multimodal models show more robustness across these specific object super categories need more exploration. One area of exploration could be the open-vocabulary training paradigm of the multimodal models where these models have been exposed to a broader and more diverse set of labels and descriptions of the aforementioned categories, enabling them to generalize better across various contexts and conditions. This wide exposure helps the models to learn more robust and transferable features that can be effective across different categories, even under distortions or corruptions. Nevertheless, this area needs a lot more exploration to understand the proper reasons behind this higher performance in certain object categories.

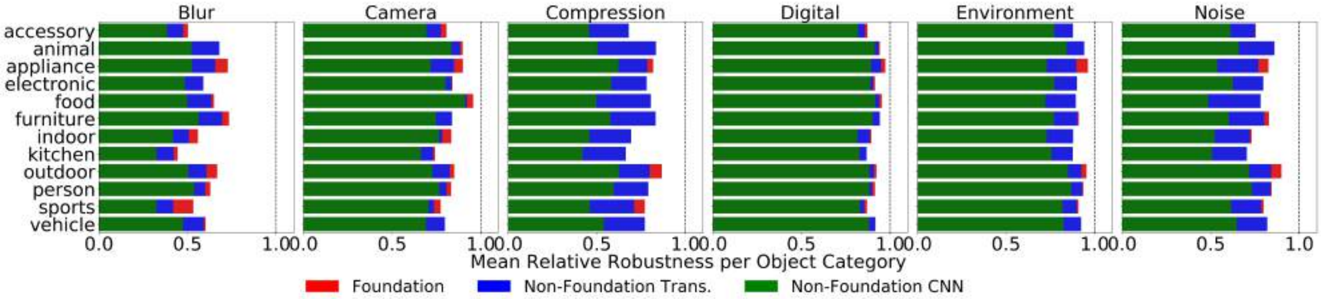

Figure 8. **Relative robustness scores $\gamma^r$ across object super categories and corruptions** categories on MS COCO-P for multimodal models and unimodal models. The higher $\gamma^r$ the more relatively robust.

Table 4. **Mean Relative Robustness ($\gamma^r$) scores for object super categories** for MS COCO-P dataset. A higher $\gamma^r$ score is more robust with the top score in bold and second underlined.

| | accessory | animal | appliance | electronic | food | furniture | indoor | kitchen | outdoor | person | sports | vehicle |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Unimodal CNN | 0.62 | 0.71 | 0.65 | 0.68 | 0.65 | 0.68 | 0.61 | 0.58 | 0.72 | 0.74 | 0.63 | 0.68 |
| Unimodal Trans. | 0.74 | **0.86** | 0.81 | **0.80** | **0.82** | 0.83 | 0.74 | **0.71** | 0.82 | 0.82 | 0.74 | **0.81** |
| Multimodal | **0.74** | 0.83 | **0.87** | 0.77 | 0.81 | **0.84** | **0.76** | 0.71 | **0.86** | **0.82** | **0.78** | 0.80 |

**The more similar corrupted image features are to original, likely more robust:** Figure 9 shows TSNE visualizations of feature spaces for image encoders from multimodal model ODISE, GroundedSAM and unimodal Mask2Former. This helps us observe whether models encode an image in the same space as its corrupted versions, clustering by image, or if it clusters by corruption type. When observing at MS COCO-P, we see that the unimodal model, Mask2Former, clusters representations by image as indicated by the overlap of different corruptions that are close to the original image. GroundedSAM seems to also be clustering by image. ODISE, on the other hand, does some clustering based on the image, but with more noticeable clustering by corruption type.

When evaluated on ADE20K-P datasets, the clustering tendency of GroundedSAM and ODISE remained the same. However, for Mask2Former, the tight clustering based on images that was observed in case of MS COCO-P is no longer present and in this case the clustering is more often based on corruption type. This aligns with robustness as well, Mask2former is more robust typically on MS COCO-P while noticeably less robust on ADE20K-P (Table 2, Table 3). This may indicate that an additional measurement of robustness is the more similar corrupted versions of an image are to the original in latent space, the more robust.

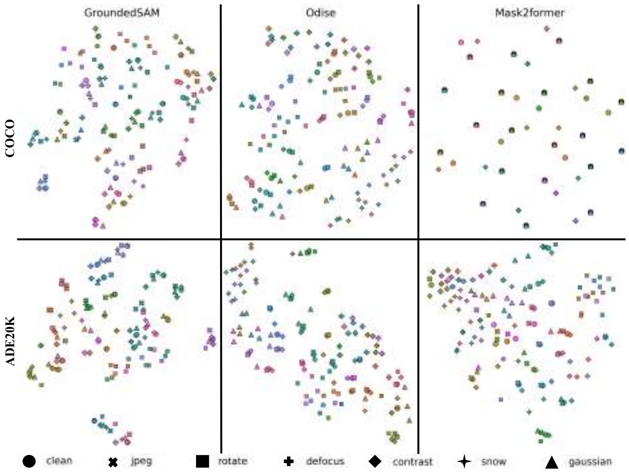

Figure 9. **Visualization of feature spaces of image encoders** from multimodal ODISE and GroundedSAM and unimodal Mask2Former. For a subset of images in MS COCO-P, we extract multiple variations under different corruptions at severity 3. Each color is a single image, while marker shape is corruption type.

## 5. Conclusion

In this benchmark, we evaluated multimodal Visual Foundation Models (VFMs) and unimodal models for segmentation on MS COCO-P and ADE20K-P datasets which are perturbed using 17 categories of corruption that reflect real-world data corruptions across 5 different level of severity. Our study provides several interesting insights about the selected models. (1) All selected models struggle with blur and and compression based corruptions (2) Although multimodal VFMs are not noticeably more robust than unimodal models however, they show competitive robustness results when evaluated zero-shot. (3) selected multimodal VFMs show higher relative robustness for specific object-types like those found in sports, outdoor and appliance compared to other unimodal models. We hope these findings and the benchmark in this work can potentially open up interesting questions about robustness segmentation and foundation models.

## References

[1] Imagebind: One embedding space to bind them all. In *CVPR*, pages 15180–15190, 2023. 2

[2] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *ICCV*, pages 10231–10241, 2021. 2

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. 3

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 1

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 2, 3, 4

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2017. 1

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. pages 1597–1607. PMLR, 2020. 1

[8] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 2, 3, 4

[9] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. 2021. 4

[10] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 2, 3, 4

[11] Dorin Comaniciu and Peter Meer. Robust analysis of feature spaces: Color image segmentation. In *CVPR*, pages 750–755. IEEE, 1997. 2

[12] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 5

[13] Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W Remedios, Shunxing Bao, Bennett A Landman, Lee E Wheless, Lori A Coburn, Keith T Wilson, et al. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. *arXiv preprint arXiv:2304.04155*, 2023. 2

[14] 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. 1

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3, 4

[16] Yunhao Ge, Jie Ren, Andrew Gallagher, Yuxiao Wang, Ming-Hsuan Yang, Hartwig Adam, Laurent Itti, Balaji Lakshminarayanan, and Jiaping Zhao. Improving zero-shot generalization and robustness of multi-modal models. In *CVPR*, pages 11093–11101, 2023. 2

[17] Yunhao Ge, Jie Ren, Jiaping Zhao, Kaifeng Chen, Andrew Gallagher, Laurent Itti, and Balaji Lakshminarayanan. Building one-class detector for anything: Open-vocabulary zero-shot ood detection using text-image models. *arXiv preprint arXiv:2305.17207*, 2023. 2

[18] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 3, 7

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

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 1

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 1, 3, 4

[22] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 2, 3

[23] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *NeurIPS*. Curran Associates, Inc., 2019. 3

[24] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 2, 3

[25] Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. imgaug. https://github.com/aleju/imgaug, 2020. Online; accessed 01-Feb-2020. 3, 1

[26] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In *CVPR*, pages 8828–8838, 2020. 2, 3, 7

[27] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models with respect to common corruptions. *IJCV*, 129:462–483, 2021. 2, 3

[28] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, pages 9404–9413, 2019. 1

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 3, 4

[30] Feng Li, Hao Zhang, Huaizhe xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *CVPR*, 2023. 1, 2, 3, 4

[31] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, pages 10965–10975, 2022. 1

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 2, 4

[33] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 3, 4

[34] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1

[35] 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 *ICCV*, 2021. 4

[36] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1

[37] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 2

[38] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 3

[39] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 4

[40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[41] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, pages 2085–2094, 2021. 2

[42] Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Philip Torr, Zhe Lin, and Jiaya Jia. Open world entity segmentation. *IEEE TPAMI*, 2022. 1

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. pages 8748–8763. PMLR, 2021. 2, 3, 4

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. pages 8748–8763. PMLR, 2021. 1

[45] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 1, 3

[46] RockeyCoss. Prompt-segment-anything. GitHub repository, 2023. 3, 4

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3

[48] Madeline Schiappa, Shruti Vyas, Hamid Palangi, Yogesh Rawat, and Vibhav Vineet. Robustness analysis of video-language models against visual and language perturbations. In *NeurIPS*, pages 34405–34420. Curran Associates, Inc., 2022. 2, 3

[49] Madeline Chantry Schiappa, Naman Biyani, Prudvi Kamtam, Shruti Vyas, Hamid Palangi, Vibhav Vineet, and Yogesh S. Rawat. A large-scale robustness analysis of video action recognition models. In *CVPR*, pages 14698–14708, 2023. 2, 3

[50] Madeline Chantry Schiappa, Naman Biyani, Prudvi Kamtam, Shruti Vyas, Hamid Palangi, Vibhav Vineet, and Yogesh S Rawat. A large-scale robustness analysis of video action recognition models. In *CVPR*, pages 14698–14708, 2023. 2, 3, 4

[51] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *NeurIPS*, 33:18583–18599, 2020. 2, 3

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 1

[53] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *CVPR*, pages 14408–14419, 2023. 1, 2, 3, 4

[54] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, pages 6830–6839, 2023. 2, 3, 4

[55] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5

[56] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023. 1, 2, 3, 4

[57] Jianwei Yang, Chunyuan Li, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Focal modulation networks, 2022. 3, 4

[58] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos, 2023. 2

[59] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *NeurIPS*, 32, 2019. 3

[60] Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. Inserting anybody in diffusion models via celeb basis. *arXiv preprint arXiv:2306.00926*, 2023. 2

[61] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 1

[62] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianfeng Gao, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. *arXiv preprint arXiv:2303.08131*, 2023. 1

[63] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015. 1

[64] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 2, 4

[65] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023. 2