Universiteit Leiden

# Master Computer Science

## OASIC: A General Framework for Occlusion-Agnostic and Severity-Informed Classification

| | |
|---|---|
| Name: | Kay Gijzen |
| Student ID: | 3724808 |
| Date: | 15/10/2025 |
| Specialisation: | Artificial Intelligence |
| 1st supervisor: | Gertjan J. Burghouts |
| 2nd supervisor: | Daniël M. Pelt |

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

# OASIC: A General Framework for Occlusion-Agnostic and Severity-Informed Classification

Kay Gijzen[1,2], Gertjan J. Burghouts[2], and Daniël M. Pelt[1]

[1] Leiden University
[2] TNO

**Abstract.** Occlusion, the partial visual obstruction of objects in an image, poses a significant challenge for state-of-the-art computer vision models. Because these models are typically trained on unoccluded data, they struggle to handle the challenges introduced by occlusion, namely the loss of visible information and the presence of distracting patterns that can mislead classifiers. To address these challenges, we propose OASIC, a method designed to enhance model robustness against both the information loss and visual distraction caused by occlusion. This thesis investigates OASIC's effectiveness in improving the resilience of existing vision models under varying occlusion levels. Our approach leverages pixel-wise occlusion likelihoods, which can be obtained from any suitable source. In this work we employ AnomalyDINO [2] to estimate the occlusion likelihoods by detecting irregularities in the image. This makes the method occlusion-agnostic, independent of the specific type or appearance of the occluder. These likelihoods are utilized in two complementary ways. First, they are used to segment and mask occluded regions with a neutral gray color, reducing visual noise. Second, they are aggregated to estimate occlusion severity, enabling severity-informed model selection that dynamically adapts to the amount of visible information. Experimental results show that combining gray masking with severity-aware model selection improves $\text{AUC}_{\text{occ}}$ by +6.36 over training directly on occluded data and by +16.65 over fine-tuning on unoccluded data. Overall, OASIC demonstrates substantial gains in occlusion robustness through a modular and adaptable framework.

**Keywords:** Computer vision · Occlusion Robustness · Visual recognition

## 1 Introduction

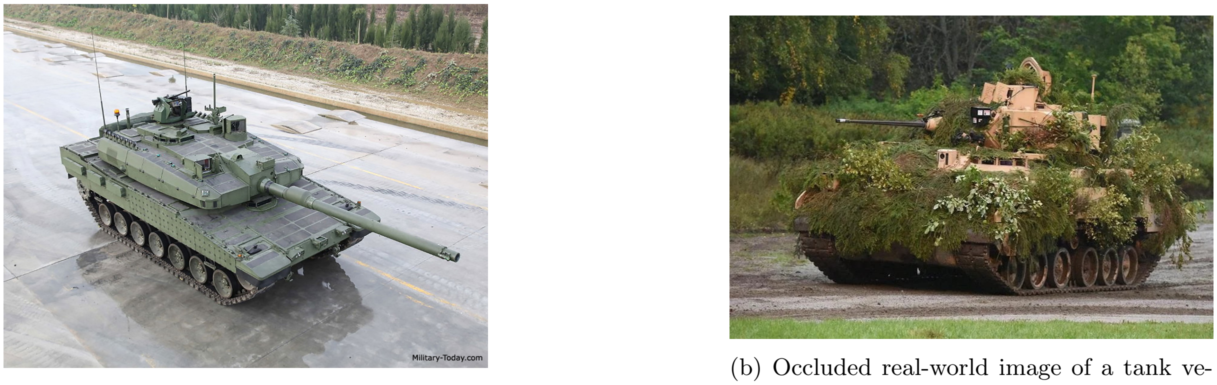
Modern computer vision models perform impressively on clean, fully visible images. In practice, however, objects are often partially hidden or obscured (for example by foliage, smoke, or other objects) causing standard models to fail. Understanding and handling such conditions is essential for deploying vision systems in real-world applications like surveillance, autonomous driving or military imagery, where occlusions are the norm rather than the exception. In this work, we focus on the latter domain, namely the fine-grained classification of military tank vehicles.

Standard datasets typically consist of fully visible examples. For instance, in military tank classification, datasets show clear, unobstructed tanks (Fig. 1a). In contrast, real-world images often capture tanks partially hidden behind terrain, foliage, or camouflage (Fig. 1b). Occlusion, the phenomenon where parts of an object are hidden from view, makes classification significantly harder.

The gap between clean training images and occluded real-world conditions is a significant challenge in computer vision [25, 27]. Occlusion not only reduces the visible regions of an object but also introduces distracting patterns that can mislead classifiers [36, 26, 30]. Occlusion handling remains an active area of research. Furthermore, datasets containing annotated occlusion are scarce, and real-world occlusions are often unpredictable, making robust fine-grained classification under occlusion a particularly difficult problem.

To address the challenges posed by occlusion, we propose *OASIC* (Occlusion-Agnostic Severity-Informed Classification), a method that explicitly targets both underlying difficulties: (i) the loss of visible information and (ii) the visual distraction introduced by occluders. Our approach leverages pixel-level maps that represent the likelihood of occlusion for each pixel. These maps can, in principle, originate from any source. However, in this work, we use *AnomalyDINO* [2], a recent patch-based anomaly detection method, to demonstrate the feasibility of our approach. We treat occluded regions as a form of visual irregularity, assuming that they deviate from the object's expected appearance and can therefore be localized as irregularities in the visual input.

By interpreting occlusion as an irregularity in the visual input, OASIC can localize *any* type of occlusion (whether vegetation, smoke, or other visual obstructions) without requiring prior knowledge of the occluder.



(a) Clean training image of a tank vehicle

(b) Occluded real-world image of a tank vehicle

Fig. 1: **Comparison of clean vs. occluded images.**

We refer to this as *occlusion-agnostic* segmentation. From the per-pixel likelihood of occlusion, we derive segmentation masks that serve two complementary purposes.

First, the segmentation is used to mitigate visual distraction: occluded regions are replaced with a neutral gray tone, effectively suppressing the misleading texture cues introduced by complex occluders. Second, we use the same segmentation maps to estimate the *occlusion severity*, defined as the fraction of the image identified as occluded. This severity estimate is then used to guide model selection. Specifically, we fine-tune a series of models on gray-occluded training images at varying occlusion severities, ensuring that training conditions align closely with those encountered at test time. During inference, the model whose training severity best matches the estimated occlusion level is selected for classification.

Fine-tuning on occluded images, however, is non-trivial. We observe that a model optimized for a single occlusion severity tends to perform poorly at others, likely because it adapts to the specific occlusion distribution present during training. As a result, robustness to one severity level does not generalize across the entire spectrum. The optimal performance at any given occlusion level depends strongly on the severity distribution the model was exposed to during training.

Together, these two components, gray masking and severity-informed model selection, form *OASIC*, a unified occlusion-aware framework designed to improve fine-grained classification robustness under varying degrees of occlusion.

### 1.1 Research questions

In this work, we address three central research questions related to understanding and mitigating the effects of occlusion in fine-grained image classification. First, we investigate how occlusion specifically impacts classification performance and model attention. Second, we examine whether AnomalyDINO can reliably segment and quantify occlusions, and how its performance compares to that of a dedicated segmentation model. Finally, we explore how the resulting occlusion information, both in terms of localization and severity, can be leveraged to enhance classification robustness under varying degrees of occlusion. Our research questions are as follows:

**RQ1**: In what ways does occlusion impact the reliability of image classification systems, considering both the reduction of observable object regions and the presence of distracting visual interference?
**RQ2**: To what extent can AnomalyDINO provide reliable occlusion segmentation and severity estimation?
**RQ3**: How can per-pixel occlusion likelihoods be integrated, through methods such as occlusion masking or severity-guided model selection, to improve fine-grained classification performance under occlusion?

### 1.2 Contributions

The main contributions of this work are as follows:

- **Occlusion-agnostic segmentation.** We introduce a method that interprets occluded regions as visual irregularities, enabling segmentation of arbitrary occluders from per-pixel anomaly likelihoods without prior knowledge of the occlusion type.

- **Gray masking and severity estimation.** Using the derived occlusion maps, we mitigate visual distraction by masking occluded regions with a neutral gray tone and estimate occlusion severity as the fraction of the image identified as occluded.
- **Severity-informed model selection.** We propose a model selection strategy guided by the estimated occlusion severity. By fine-tuning models across multiple occlusion levels, our approach adaptively selects the most suitable model at inference, leading to improved classification robustness under occlusion.

### 1.3 Thesis structure

The remainder of this thesis is structured as follows. Section 2 reviews related work and introduces the necessary preliminaries. Section 3 describes the dataset used in this study. Section 4 presents our proposed method, OASIC. The experimental setup is detailed in Section 5, followed by the results in Section 6. Finally, Section 7 discusses the findings and their implications.

## 2 Related work and preliminaries

### 2.1 Occlusion handling

Occlusion poses a major challenge in visual recognition because it hides informative regions and introduces misleading signals. To address this, researchers have proposed a variety of strategies aimed at improving robustness. Broadly, these methods fall into two categories. The first category leverages data augmentation techniques, which expose models to occlusion-like scenarios during training so that they learn to rely on multiple image regions rather than a single discriminative patch. The second category focuses on part-based modeling, where objects are broken down into semantic parts, so the model can still make predictions even if only a subset of those parts is visible. Together, these approaches reflect the current research trend of reducing the impact of missing or corrupted information in occluded images.

**Data augmentation techniques** To handle occlusion, researchers have introduced data augmentation techniques that create modified training samples, discouraging models from relying too strongly on any single image region for predictions. Mixup [35] creates new training samples by blending two images together through a weighted average of their pixels. CutMix [33] instead replaces regions of one image with patches from another, producing cut-and-paste combinations. Hide-and-Seek [15] randomly hides patches in an image so the model is forced to look at other useful regions when the most obvious ones are missing. The common idea across these methods is that by training on partial or mixed views of images, networks are encouraged to learn part-based features. TransMix [1], an extension of CutMix, goes a step further by adjusting how much of each image contributes to the final label based on transformer attention values. The authors argue this better captures part-based knowledge, which helps models perform more reliably under occlusion. In general, data augmentations like Mixup, CutMix and TransMix can improve recognition on occluded images because they splice together image parts and link labels to visible regions. However, while these methods reduce the loss of information caused by occlusion, they do not fully address the confusion introduced when irrelevant or misleading features are present.

**Part-based modeling** Part-based methods break objects into semantic parts, classify those parts, and then combine them to predict the full object. The idea is that even if part of an object is hidden, the visible parts still provide useful information. CompositionalNets [14, 13] include an occlusion localization module that predicts which regions are blocked. Features from a standard convolutional neural network (CNN) backbone are organized into dictionaries that act like part detectors, capturing how object parts are arranged across classes. By reasoning explicitly over part visibility, CompositionalNets improve recognition when objects are partially occluded. TDMPNet [31] also estimates a visibility map from CNN features, but it suppresses irrelevant or occluded features earlier in the network. Its top-down attention module removes activations caused by occluders, producing cleaner feature representations. By filtering out noise, TDMPNet helps the model focus only on visible object parts and reduces errors caused by partial occlusion.

While these methods show the promise of part-based reasoning, they are still limited by their CNN backbones. CNNs are limited by their local receptive fields, which makes it difficult for them to capture long-range dependencies and reason about object parts that are far apart or partially hidden. Additionally, research has found that CNNs are not robust to occlusion [5], and are prone to overfitting, which in turn leads to poor generalization when occlusion is present [3].

### 2.2 Vision transformers

The Vision Transformer (ViT) adapts the transformer architecture [29] to images and has shown strong robustness to occlusion. Unlike CNNs, which build larger receptive fields layer by layer, ViTs split an image into patches and model their global relationships through self-attention [4]. This global view allows ViTs to recover context even when large areas of an image are hidden.

Studies show that ViTs outperform CNNs under occlusion [17, 9]. A key factor is self-supervised pretraining, especially Masked Image Modeling (MIM) [12, 32]. Similar to masked language modeling in natural language processing, MIM hides large portions of the input and trains the network to reconstruct them. Approaches such as Masked Autoencoders [8] and iBOT [37] build on this principle, making ViTs somewhat resilient to occlusion-like scenarios. Large-scale foundation models such as CLIP [21] and DINOv2 [18] use the transformer architecture introduced in ViTs and are pretrained on massive, diverse datasets. This pretraining granted them with rich feature representations and broad semantic knowledge, allowing them to act as general-purpose vision backbones that can be adapted to a wide range of tasks with little or no fine-tuning.

Unlike augmentation- and part-based methods, which are designed explicitly to handle occlusion, transformer-based models gain robustness more indirectly through their pretraining strategies. This incidental robustness has been observed in practice: transformer-based models retain competitive accuracy even when large portions of the input are occluded [17]. However, this robustness is uneven and breaks down under stronger or structured occlusions, such as when key object parts are consistently hidden [9]. This suggests that complementary strategies may be necessary to move beyond incidental resilience and achieve reliable performance under occlusion.

### 2.3 Occlusion segmentation

To the best of our knowledge, no prior work has addressed occlusion segmentation in an *occlusion-agnostic* manner by leveraging pixel-level anomaly maps. In this work, we repurpose anomaly detection techniques—specifically *AnomalyDINO* [2] (which was originally developed for defect localization) as a generalized occlusion estimator. Segmenting "the unusual" is a common paradigm in industrial anomaly detection, where methods such as AnomalyDINO, PatchCore [23] and DRAEM [34] generate per-pixel anomaly likelihoods to localize defects or irregularities.

Our approach extends this principle beyond industrial inspection: we interpret high anomaly likelihoods as indicators of occluded regions. This enables occlusion segmentation without any explicit occlusion supervision, relying solely on image-level labels. While no existing work has yet generalized occlusion segmentation in this way, related segmentation methods can target specific occluder types. For instance, SAM2 [22], a state-of-the-art zero-shot segmentation method, and OVSeg [16], an open-vocabulary segmentation model, can be prompted to segment known occluders such as vegetation [28]. However, such approaches require prior knowledge of the occluder type at inference time, along with model prompting or specialization for each occlusion category. This is an inherent limitation our occlusion-agnostic formulation overcomes.

### 2.4 Preliminary: AnomalyDINO

To quantify occlusions at the pixel level, we employ **AnomalyDINO** [2]. AnomalyDINO is a method originally designed to produce anomaly maps indicating the probability of each pixel being anomalous. In our case, these anomaly maps are interpreted as occlusion probability maps. AnomalyDINO is a recent anomaly detection method that builds on the strong patch-level features extracted by DINOv2 [18]. AnomalyDINO is a *training-free, patch-based technique*, which makes it especially suitable for few-shot scenarios and for settings where collecting large-scale anomaly data is impractical.

**Patch-level nearest-neighbor scoring** AnomalyDINO operates as a nearest-neighbor approach applied to patch-level feature embeddings. Each image is represented as a collection of *patch features*, which are compared to a memory bank of nominal reference patches to compute per-patch anomaly scores. This highlights regions that deviate from expected appearance, such as occluded areas, which receive higher anomaly scores.

The memory bank $\mathcal{M}$ is constructed by extracting patch-level embeddings from a set of nominal (non-occluded) reference images using DINOv2. For a test image, patch embeddings are computed in the same manner and compared against $\mathcal{M}$. The anomaly score of a test patch $\mathbf{p}$ is defined as its nearest-neighbor distance to the reference patches:

$$d_{\mathsf{NN}}(\mathbf{p};\mathcal{M}) = \min_{\mathbf{p}_{\mathsf{ref}}\in\mathcal{M}} d(\mathbf{p},\mathbf{p}_{\mathsf{ref}}), \tag{1}$$

where $d(\cdot,\cdot)$ denotes a distance metric. Following the original AnomalyDINO paper, we use the cosine distance, defined as

$$d_{\mathsf{cos}}(\mathbf{x},\mathbf{y}) = 1 - \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|}, \tag{2}$$

where $\mathbf{x}\cdot\mathbf{y}$ is the dot product of vectors $\mathbf{x}$ and $\mathbf{y}$, and $\|\mathbf{x}\|$ denotes the Euclidean norm of $\mathbf{x}$. Intuitively, if a patch in the test image does not resemble any nominal reference patch, its anomaly score will be high.

**Continuous anomaly map** After computing patch-level anomaly scores $d_{\mathsf{NN}}(\mathbf{p};\mathcal{M})$ for each patch $\mathbf{p}$, we transform these discrete patch scores into a *continuous per-pixel anomaly map* $A \in [0,1]^{H\times W}$, where $H$ and $W$ denote the height and width of the image. Let $D \in \mathbb{R}^{h\times w}$ denote the matrix of patch-level distances, normalized to $[0,1]$. The continuous anomaly map is obtained by applying bilinear upsampling $\mathcal{U}$ followed by Gaussian smoothing $\mathcal{G}_\sigma$:

$$A = \mathcal{G}_\sigma\Big(\mathcal{U}(D)\Big), \quad \sigma = 4.0, \tag{3}$$

where $\mathcal{U} : \mathbb{R}^{h\times w} \to \mathbb{R}^{H\times W}$ performs bilinear interpolation, and $\mathcal{G}_\sigma$ denotes Gaussian smoothing with standard deviation $\sigma$. Each entry of $A$ satisfies $0 \leq A_{i,j} \leq 1$, where $0$ indicates a non-anomalous (normal) pixel and $1$ indicates a highly anomalous (potentially occluded) pixel.

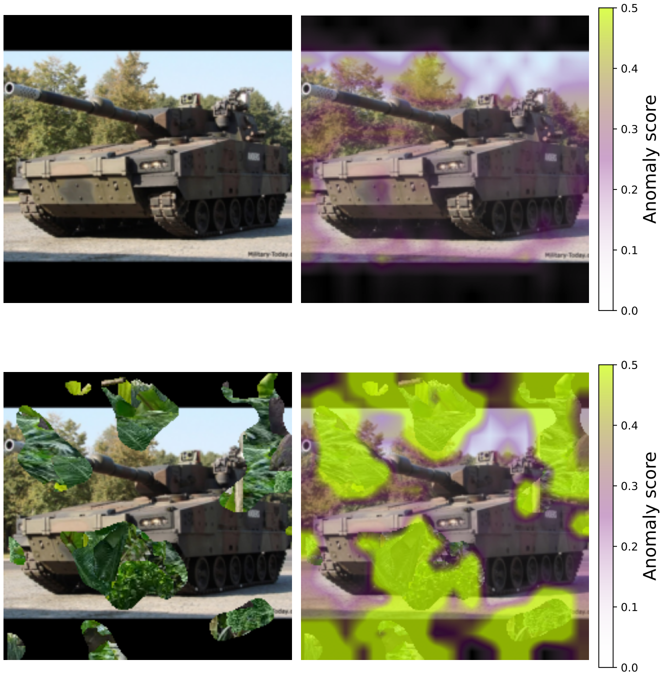


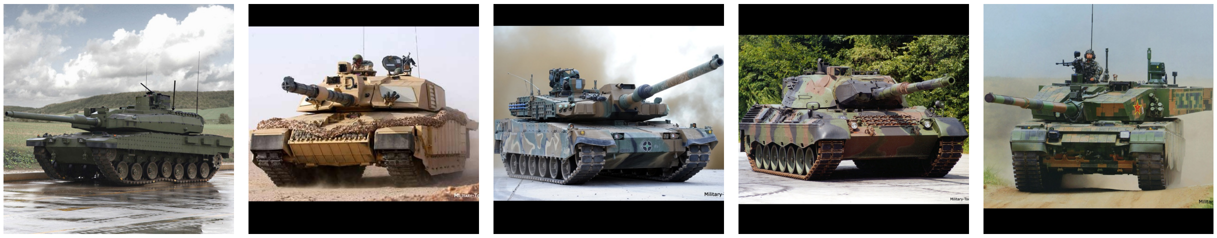
Fig. 2: **Visualization of anomaly detection.** The top row shows the anomaly maps overlaid on a clean image, highlighting potential anomalies. The bottom row shows the anomaly map overlaid on an occluded image, demonstrating how occlusion affects the anomaly detection results.

As shown in Figure 2, the top row shows the anomaly map overlaid on a clean image, highlighting the regions identified as anomalous. The bottom row presents the corresponding anomaly map on an occluded image, demonstrating how occlusions influence the anomaly detection results. This comparison clearly shows the effect of occlusions on the anomaly predictions.

# 3 Dataset

While dataset descriptions are typically included in the experimental setup, we present it earlier here because the artificial occlusion generation process introduced below is a prerequisite for understanding our proposed method.

We conduct our experiments on a fine-grained, non-occluded dataset of military tank images, comprising 205 training images and 41 test images across 29 classes. Each class corresponds to a specific type of tank vehicle, resulting in subtle inter-class variations. Accurate classification therefore depends on recognizing small visual differences, making effective occlusion handling particularly important in this context.Figure 3 illustrates five examples from five distinct tank classes, highlighting the subtle but critical visual variations that distinguish them.



(a) Altay (b) Challenger 2 (c) K2 Black Panther (d) Leopard (e) Type 99

Fig. 3: **Five example images from distinct tank classes in the dataset, illustrating subtle inter-class visual differences.**

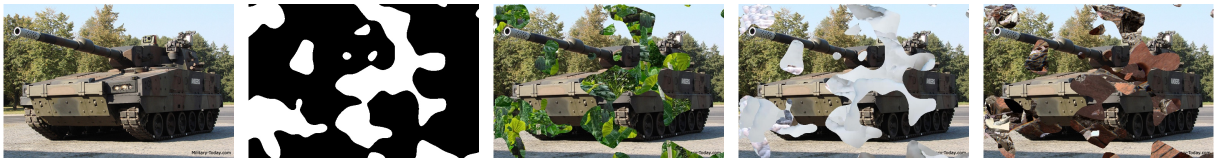
It is unlikely that foundation models such as DINOv2 have encountered our dataset during pretraining. This makes it particularly relevant for our study, as we fine-tune the model rather than relying solely on pretrained representations, ensuring that our experiments remain largely independent of the original pretraining data.

Although the base dataset is unoccluded, it can be scaled to arbitrary occlusion levels by applying artificial occlusion during training and evaluation. Rather than precomputing occluded versions, we generate occlusions on the fly: for each image, an occlusion mask is procedurally created and immediately applied by overlaying cutouts of occluding objects. This approach enables control over both the type and severity of occlusion during training.

## 3.1 Artificial occlusion generation

To study the effect of occlusion and evaluate our approach, we generate artificial occlusions on the non-occluded images. First, we create ground-truth occlusion masks $M$ using Perlin noise [20], a gradient noise function widely used in computer graphics to produce natural textures such as clouds, smoke, and terrain. Its spatial coherence ensures smooth patterns, making it well-suited for simulating gradually varying occlusions like vegetation, smoke or fog. Importantly, this approach allows us to actively control the degree of occlusion, i.e., the percentage of pixels in an image that are covered by the occluding region. The occlusion masks provide pixel-level ground truth, enabling quantitative evaluation of occlusion segmentation and severity estimation. An example of the artificial occlusion generation can be seen in Figure 4.

Using these masks, we overlay occluding objects onto the original images to generate occluded samples. We first build a bank of occlusion cutouts by segmenting natural images containing the desired occlusion objects (for example foliage, smoke, or rubble) using the Segment Anything Model (SAM) [10]. The resulting cutouts are saved and later reused during training and evaluation, where they are then pasted inside the regions of the occlusion mask. This approach allows us to apply diverse occlusions on-the-fly while maintaining consistency across experiments. This procedure yields controlled, realistic occlusion scenarios, effectively enriching the dataset for evaluating occlusion robustness.



(a) Original sample (b) Generated occlusion mask (c) Sample with vegetation occlusion (d) Sample with smoke occlusion (e) Sample with rubble occlusion

Fig. 4: **Examples of occlusion generation.** (a) original dataset image, (b) Perlin-generated occlusion mask, used to overlay the clean image with a textured occlusion, resulting in (c) vegetation-occluded image, (d) smoke-occluded image, and (e) rubble-occluded image.

We refer to any artificial occlusion introduced through cutouts, such as vegetation, smoke, or rubble, as a textured occlusion. In addition, we consider a non-textured occlusion, obtained by overlaying a uniform gray mask on the image.

## 4 OASIC: Occlusion-agnostic severity-informed classification

In this section, we present our proposed method, **Occlusion-Agnostic Severity-Informed Classification (OASIC)**. We begin with a high-level overview of the approach, followed by detailed explanations of its main components: occlusion map generation, occlusion segmentation, severity estimation, and model selection.

### 4.1 Overview

We propose OASIC (Fig. 5), a framework that tackles the two primary challenges introduced by occlusion: visual distraction caused by misleading occlusion textures, and information loss due to missing object regions. To address these issues, our approach leverages pixel-level maps that estimate per-pixel occlusion likelihoods and uses them in two complementary ways.

First, we convert pixel-wise occlusion likelihoods into occlusion maps that localize regions likely affected by occlusion. These regions are then masked with a neutral gray, replacing distracting occlusion textures with a uniform appearance. This step directly mitigates the effect of visual distraction, ensuring that the model focuses on visible and relevant object features while preserving the overall image structure.

Second, by aggregating pixel-wise occlusion likelihoods, we estimate the occlusion severity, expressed as the fraction of the image that is occluded. This estimate serves as a proxy for the available visual information and guides the selection of the most suitable classifier from a pool of models fine-tuned for different occlusion severities. In doing so, the framework adapts to varying levels of information loss, ensuring stable recognition performance across diverse visibility conditions. Our experiments further confirm that no single model performs optimally across all occlusion severities, highlighting the need for such severity-informed model selection.

Our approach leverages occlusion likelihoods to mitigate distraction from the primary object and to adapt recognition under reduced object visibility. In this work, we employ **AnomalyDINO** [2], an occlusion-agnostic method that generates anomaly maps indicating the likelihood of each pixel being anomalous. The method is considered occlusion-agnostic because it detects irregularities in appearance rather than relying on predefined occluder types (e.g., vegetation or other domain-specific artifacts). We interpret these anomaly maps as occlusion probability maps, reflecting how likely each pixel is to belong to an occluding region.

Although AnomalyDINO is used in our implementation, the proposed framework remains independent of the specific source of these maps. Any method capable of producing pixel-level likelihoods of occlusion could be integrated. For simplicity of notation, we refer to these pixel-level likelihood maps, regardless of their origin, as anomaly maps throughout this thesis.

### 4.2 Occlusion segmentation and masking

AnomalyDINO produces an anomaly map $A \in [0,1]^{H \times W}$ that assigns each pixel a likelihood of being occluded. To obtain a discrete representation of these regions, the map is thresholded at a value $\tau$, yielding a binary occlusion map $O \in \{0,1\}^{H \times W}$ that indicates which pixels are classified as occluded:

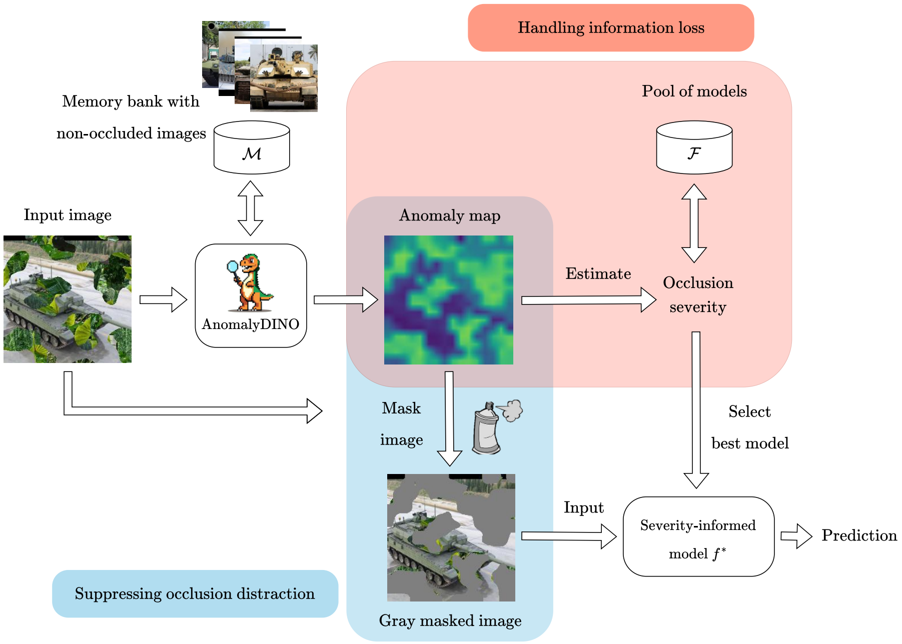


Fig. 5: **Occlusion handling with OASIC.** Unoccluded representations of the to-be-classified objects are collected and stored in the memory bank $\mathcal{M}$, while a pool of models $\mathcal{F}$ is fine-tuned for optimal performance under varying occlusion severities. At test time, an anomaly map is inferred using *Anomaly-DINO* by scoring against the memory bank. The anomaly map is then used to both segment and quantify occlusions: the segmented masks guide gray masking of occluded regions (to suppress distraction), while the estimated severity informs the selection of the most suitable classification model $f^*$ from the pool $\mathcal{F}$ (to better handle reduced visual information). Finally, classification is performed on the gray-masked image using $f^*$.

$$O_{i,j} = \begin{cases} 1, & \text{if } A_{i,j} \geq \tau, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where $\tau \in [0, 1]$. The choice of threshold $\tau$ allows control over the aggressiveness of occlusion detection:

- *Aggressive detection:* Use a low threshold to capture as much occlusion as possible, at the cost of potential false positives.
- *Conservative detection:* Use a high threshold to mark only high-confidence occluded pixels, reducing false positives.

**Adaptive thresholding** Instead of using a fixed threshold $\tau$, we also explore dynamic thresholding based on Otsu's method [19]. This approach selects $\tau$ by analyzing the histogram of anomaly scores in $A$ and maximizing the between-class variance, thus adapting the thresholded binary occlusion map $O$ to each image. Let the normalized histogram of anomaly values be $p(i)$ for intensity levels $i \in \{0, \ldots, L-1\}$. For a given threshold $t$, the class probabilities and class means are defined as:

$$\omega_0(t) = \sum_{i=0}^{t} p(i), \qquad \omega_1(t) = \sum_{i=t+1}^{L-1} p(i), \tag{5}$$

$$\mu_0(t) = \frac{1}{\omega_0(t)} \sum_{i=0}^{t} i\, p(i), \qquad \mu_1(t) = \frac{1}{\omega_1(t)} \sum_{i=t+1}^{L-1} i\, p(i). \tag{6}$$

The between-class variance is then given by:

$$\sigma_b^2(t) = \omega_0(t)\,\omega_1(t)\,[\mu_0(t) - \mu_1(t)]^2. \tag{7}$$

Otsu's method selects the optimal threshold as:

$$\tau^* = \arg\max_t \sigma_b^2(t). \tag{8}$$

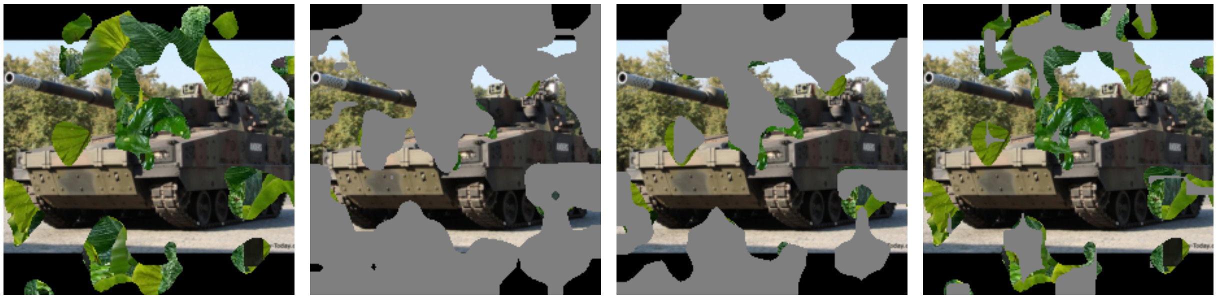
Finally, the binary occlusion map $O$ is obtained as:

$$O_{i,j} = \begin{cases} 1, & \text{if } A_{i,j} \geq \tau^*, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

**Occlusion masking** Using the binary occlusion map $O$, we mask occluded regions with a uniform gray value. This suppresses the high-frequency textures typically caused by occlusions, preventing them from interfering with feature extraction and classification. Specifically, we construct a masked image $I_{\text{masked}}$ by replacing all pixels identified as occluded ($O_{i,j} = 1$) with a uniform gray value $g$:

$$I_{\text{masked},i,j} = \begin{cases} g, & \text{if } O_{i,j} = 1, \\ I_{i,j}, & \text{otherwise,} \end{cases} \tag{10}$$

where $I$ denotes the original image and $g$ denotes the constant gray intensity applied to occluded pixels. Considering we use RGB images, we set $g = 127$ for all channels, corresponding to a mid-level gray tone. This procedure removes distracting appearance artifacts introduced by occlusion, while preserving the visible, non-occluded regions of the object. The resulting masked images serve as input for downstream tasks, in our case being classification. As shown in Figure 6, the effect of different threshold values $\tau$ on the masking is clearly visible.



(a) Occluded image (b) Masked image, $\tau = 0.3$ (c) Masked image, $\tau = 0.5$ (d) Masked image, $\tau = 0.7$

Fig. 6: **Comparison of an occluded image and its masked versions at different thresholds $\tau$.** From left to right: the original occluded image, and masks applied with thresholds 0.3, 0.5, and 0.7.

### 4.3 Occlusion severity estimation

Beyond binary occlusion segmentation, we also quantify the *severity* of occlusion, defined as the proportion of the image that is occluded. Formally, we denote the estimated occlusion severity as $\hat{s} \in [0, 1]$, representing the fraction of occluded pixels in the image. We evaluate three strategies to estimate occlusion severity from the anomaly map $A$ and the binary occlusion map $O$:

- *Mean anomaly score:* The continuous anomaly map $A$ provides per-pixel anomaly scores. The severity is approximated as the spatial mean:

$$\hat{s}_{\text{mean}} = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} A_{i,j}. \tag{11}$$

- *Fixed-threshold proportion:* Given a threshold $\tau$, the binary occlusion map $O$ can be used to compute the proportion of occluded pixels:

$$\hat{s}_{\tau} = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} O_{i,j}. \tag{12}$$

- *Adaptive-threshold proportion:* Using Otsu's method to derive a dynamic threshold $\tau_{\text{Otsu}}$, we compute a binary occlusion map $O_{\text{Otsu}}$ and define severity as:

$$\hat{s}_{\text{Otsu}} = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} O_{i,j}^{\text{Otsu}}. \tag{13}$$

These three formulations allow us to compare soft (continuous) and hard (discrete) interpretations of the anomaly map, as well as fixed and adaptive thresholding strategies. In all cases, the severity value provides a measure of occlusion extent, which is used within our framework to select the most appropriate model for the given level of occlusion.

### 4.4 Fine-tuning for occlusion robustness

Empirically, we observed that models fine-tuned on a range of occlusion levels $[0, p]$ achieve their peak performance when evaluated on test images with occlusion severities close to $p$. At the same time, these models maintain competitive performance on images with lower occlusion severities ($< p$), indicating that exposure to a moderate range of occlusions promotes broader robustness.

To obtain models specialized for different levels of occlusion, we fine-tune multiple instances of the base classifier on synthetically occluded datasets. For each maximum occlusion level $p \in \{0, 10, \ldots, 100\}$, we construct a corresponding training dataset $\mathcal{D}_{[0,p]}$ by applying gray occlusions with severities uniformly sampled from the range $[0, p]$. Each image in $\mathcal{D}_{[0,p]}$ is occluded with neutral gray regions covering a random fraction $p'$ of its area, where $p' \sim \mathcal{U}(0, p)$. Fine-tuning the base classifier on $\mathcal{D}_{[0,p]}$ yields a model denoted by $f_{[0,p]}$.

This approach exposes each model to a range of occlusion severities up to $p\%$, allowing it to adapt its feature representations accordingly. Furthermore, by applying a similar gray-masking at inference (guided by the estimated occlusion map) we simulate the visual conditions encountered during testing, thereby promoting consistency between training and inference.

### 4.5 Severity-informed model selection

We observed that models fine-tuned on datasets with specific occlusion ranges (e.g., $[0, p]$) tend to perform best on test images whose occlusion severity lies near the upper bound of that range. Performance gradually degrades as the test occlusion level deviates from the training range, suggesting that a single model may not perform optimally across the full spectrum of occlusion severities. We further investigate this behavior in the Experiments section.

To address this limitation, we maintain a pool of fine-tuned models

$$\mathcal{F} = \{f_{[0,p]} \mid p \in \mathcal{P}\},$$

where $\mathcal{P}$ denotes the set of maximum occlusion levels used during fine-tuning. During inference, we estimate the occlusion severity of an input image as $\hat{s} \in [0, 1]$.

We then select the most suitable model $f_{[0,p^*]}$ from $\mathcal{F}$ based on the estimated severity $\hat{s}$:

$$p^* = \arg\min_{p \in \mathcal{P}} |\hat{s} - p|, \tag{14}$$

$$f^* = f_{[0,p^*]}. \tag{15}$$

This severity-informed selection ensures that each image is processed by the model best suited to its occlusion severity, thereby improving classification performance across varying levels of object visibility.

## 5 Experimental Setup

This section outlines the experimental setup used to evaluate the proposed method. We describe model architectures, and training configurations employed throughout our experiments. Furthermore, we detail the evaluation metrics used to assess the effectiveness of our approach in improving classification robustness under occlusion.

### 5.1 Classifier architecture and training scheme

For all experiments, we employ the same fine-grained classification model to ensure comparability across different conditions. The classifier consists of a **DINOv2** feature extractor based on a pretrained Vision Transformer (ViT-B/14), combined with a multilayer perceptron (MLP) head. We will refer to this model as the *DINOv2-MLP classifier* throughout this thesis.

The MLP head receives the 768-dimensional feature embeddings from the DINOv2 backbone and consists of a single hidden layer with 512 units, followed by a ReLU activation and a dropout layer ($p = 0.2$) to reduce overfitting. The hidden representation is then projected to the number of classes via a fully connected output layer:

$$h(x) = \mathsf{Linear}_{768\to512} \;\to\; \mathsf{ReLU} \;\to\; \mathsf{Dropout}(0.2) \;\to\; \mathsf{Linear}_{512\to C} \tag{16}$$

where $C$ is the number of classification categories.

**Fine-tuning strategy** The fine-tuning procedure is structured in stages to gradually adapt the pretrained backbone while preventing catastrophic forgetting [6]:

- **Epochs 1–5:** Only the MLP head is trained, while the DINOv2 backbone remains frozen.
- **Epochs 6–15:** The MLP head and the last three layers of the DINOv2 backbone are trained jointly.
- **Epochs 16–20:** The entire network, including all layers of DINOv2, is fine-tuned.

Depending on the experiment, the classifier is fine-tuned on images with different occlusion types (e.g., gray occlusions or textured occlusions such as vegetation). The exact fine-tuning configuration used will be specified. We

**Optimization** We use the Adam optimizer throughout training. Separate learning rates are applied to the backbone and the MLP head: a learning rate of $5 \times 10^{-5}$ is used for the DINOv2 parameters, and $5 \times 10^{-3}$ for the MLP parameters.

### 5.2 AnomalyDINO Parameters

We use AnomalyDINO to generate anomaly maps. It uses a memory bank of non-anomalous (unoccluded) reference images. The primary parameter in the AnomalyDINO method is the size of the memory bank, which specifies the number of reference images used to construct the patch feature bank $\mathcal{M}$. A larger memory bank provides more diverse reference patches, potentially improving anomaly detection, while a smaller memory bank reduces memory requirements and computational cost. In our implementation, we construct the memory bank by sampling $k$ reference images per class. The impact of the choice of $k$ on occlusion segmentation performance is evaluated and discussed in the Results section.

### 5.3 Evaluation metrics

**Occlusion robustness** We want to know how well a classifier performs under occlusion. To quantify this, we measure the classifier's accuracy under increasing levels of occlusion and summarize it using the Area Under the Curve (AUC). This metric captures how well the classifier maintains performance as occlusion increases: higher values indicate better overall robustness.

Let $p_0, p_1, \ldots, p_n$ denote the discrete occlusion levels applied to the input images, and let $\mathrm{Acc}(p_i)$ be the classification accuracy at occlusion level $p_i$. The *Area Under the Curve* (AUC) of the accuracy-under-occlusion curve, using a trapezoidal approximation over the discrete steps, is defined as:

$$\mathrm{AUC}_{\mathrm{occ}} = \frac{1}{n} \sum_{i=0}^{n-1} \frac{\mathrm{Acc}(p_i) + \mathrm{Acc}(p_{i+1})}{2}. \tag{17}$$

**Saliency consistency under occlusion** To assess how robust the classifier is under occlusion, we evaluate the consistency of its saliency maps. We use EigenGradCAM to compute a saliency map $S_{\mathsf{occ}}$ on occluded images. Since we artificially control the occlusion, we also obtain the corresponding binary occlusion mask $M_{\mathsf{occ}}$.

For reference saliency, we first compute $S_{\mathsf{clean}}$ on the clean (unoccluded) image. By masking out occluded regions, we obtain a ground-truth saliency target:

$$S_{\mathsf{gt}} = S_{\mathsf{clean}} \cdot (1 - M_{\mathsf{occ}}). \tag{18}$$

Depending on the evaluation objective, we quantify the overlap of $S_{\mathsf{occ}}$ either with the ground-truth saliency $S_{\mathsf{gt}}$ (to measure preservation of relevant attention), or directly with the occlusion mask $M_{\mathsf{occ}}$ (to measure whether the model erroneously attends to occluded regions).

Given two normalized maps $A, B \in [0, 1]$, the overlap is defined as

$$\mathsf{Overlap}(A, B) = \frac{1}{N} \sum_{i=1}^{N} \min\left(A^{(i)},\, B^{(i)}\right), \tag{19}$$

where $i$ indexes pixels and $N$ is the total number of pixels. In our experiments, we report

$$\mathsf{Overlap}_{\mathsf{GT}} = \mathsf{Overlap}(S_{\mathsf{gt}},\, S_{\mathsf{occ}}), \tag{20}$$

$$\mathsf{Overlap}_{\mathsf{Occ}} = \mathsf{Overlap}(M_{\mathsf{occ}},\, S_{\mathsf{occ}}). \tag{21}$$

$\mathsf{Overlap}_{\mathsf{GT}}$ measures the overlap between the saliency map on the clean image $S_{\mathsf{gt}}$ and the saliency map on the occluded image $S_{\mathsf{occ}}$. A higher value indicates that the model's attention is stable under occlusion, i.e., it continues to focus on the same relevant regions as in the clean image. $\mathsf{Overlap}_{\mathsf{Occ}}$ measures the overlap between the occlusion mask $M_{\mathsf{occ}}$ and the saliency map on the occluded image $S_{\mathsf{occ}}$. A lower value is better, since it means the model avoids focusing on occluded (irrelevant) areas.

**Segmentation performance** For the occlusion segmentation task, we use threshold-independent metrics that are widely used in binary segmentation. Specifically, we report the Receiver Operating Characteristic (ROC) curve with its corresponding Area Under the Curve (AUROC), as well as the Precision–Recall (PR) curve with its corresponding Average Precision (AP). These metrics evaluate segmentation performance across all possible thresholds.

# 6 Results

In this section, we present the experimental results of our study. We begin by analyzing the impact of occlusion on a fine-grained classification task. Next, we compare occlusion segmentation performance using AnomalyDINO and a baseline approach, and evaluate the effect of masking occluded regions in gray. We then explore fine-tuning strategies to improve robustness to occlusion and estimate occlusion severity. Based on this estimation, we select the best severity-informed model. Finally, we demonstrate the effectiveness of our approach on a fine-grained classification task.

## 6.1 Impact of occlusion on fine-grained classification

Occlusion inherently removes information from an image, hiding object parts that are essential for accurate predictions. However, visibility loss may not be the sole contributing factor. We hypothesize that textured occlusions (such as vegetation, smoke, or rubble) do more than merely block the view: they can actively distract the model. Their visual complexity may mislead the classifier, drawing attention toward the occluded regions instead of the object itself.

In contrast, a neutral gray occlusion should introduce less distraction, offering a cleaner "absence" rather than a competing visual signal. If this holds true, models may appear more robust when tested under gray occlusions than under textured ones.

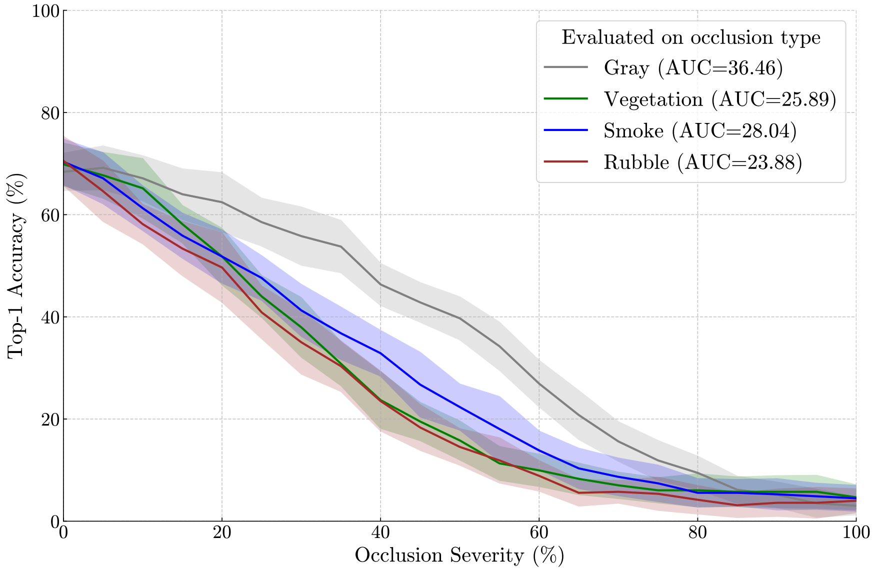


Fig. 7: **Performance of a classifier trained only on clean images, evaluated under four occlusion types.** The figure shows accuracy across occlusion severities. Performance drops smoothly under gray occlusion (gray) but more sharply for vegetation (green), smoke (blue), and rubble (brown), indicating reduced robustness to textured occlusions. The AUC quantifies occlusion robustness, with gray achieving the highest robustness compared to textured occlusions.

**Performance under occlusion** To test whether a classifier performs better under gray occlusion than under textured occlusion, we fine-tune the DINOv2-MLP classifier on clean images only and evaluate its performance under four distinct occlusion conditions: gray, vegetation, smoke, and rubble. Classification accuracy is measured across varying occlusion severities, with results shown in Figure 7. The figure also reports the corresponding $\text{AUC}_{\text{occ}}$ values, providing a compact view of performance degradation under each occlusion type.

As anticipated, the classifier achieves its highest performance under gray occlusion (gray), while textured occlusions cause a more pronounced drop in accuracy. This aligns with our expectation that uniform gray occlusion would be less disruptive, as it removes information without introducing additional visual distractions. In contrast, textured occlusions, such as vegetation or rubble, appear to mislead the model, diverting attention away from the visible object regions. Building on these findings, we next examine where the model directs its attention under occlusion, focusing on whether textured occlusions induce shifts in attention compared to uniform gray occlusions.

**Analyzing classifier attention** To analyze the classifier's attention under occlusion, we compare its focus on gray and textured occluded images. We hypothesize that the classifier struggles to maintain attention on the object. Particularly under textured occlusions, and to a lesser extent under gray ones—sometimes redirecting focus toward the occluded regions themselves. To visualize these effects, we employ **EigenGrad-CAM**[3] on the final layer of the DINOv2 feature extractor, which captures high-level semantic representations. For both clean and occluded images, we generate saliency maps, using those from the clean images as a reference baseline.

To quantify how much the classifier's attention drifts, we compute two saliency overlap scores. We use the overlap metric as defined in Section 5.3. We report two types of saliency overlap:

(i) $\text{Overlap}_{\text{GT}}$: the overlap between the clean-image saliency $S_{\text{gt}}$ and occluded-image saliency $S_{\text{occ}}$, measuring attention consistency; and
(ii) $\text{Overlap}_{\text{Occ}}$: the overlap between $S_{\text{occ}}$ and the occlusion mask $M_{\text{occ}}$, measuring how much attention leaks into occluded regions.

[3] Implemented using the `pytorch-grad-cam` library [7].

To summarize these effects across occlusion severities, we integrate the overlap scores over all tested levels, resulting in the AUC of each overlap metric.

To summarize the effects across different occlusion severities, we aggregate the saliency-overlap scores computed at $p \in \{0, 10, 20, \ldots, 100\}$ into a single statistic: the area under the Overlap at $p$ curve (AUC), giving a single AUC value per overlap metric. Intuitively, a larger AUC indicates that, across occlusion levels, the model achieves more saliency overlap. Whereas a smaller AUC reflects less saliency overlap.

| Occlusion type | AUC | |
|---|---|---|
| | $\text{Overlap}_{\text{GT}}$ (↑) | $\text{Overlap}_{\text{Occ}}$ (↓) |
| Gray | **30.33±2.27** | **37.45±2.38** |
| Vegetation | 25.43±2.01 | 47.52±1.93 |
| Smoke | 28.03±2.04 | 41.37±2.50 |
| Rubble | 26.42±1.87 | 47.20±1.93 |

Table 1: **Saliency overlap with respect to ground truth saliency (GT) and occluded regions (Occ).** Higher $\text{Overlap}_{\text{GT}}$ and lower $\text{Overlap}_{\text{Occ}}$ indicate better performance. For each metric we highlight the **best result** in bold. Gray occlusions yield the best results, with consistently higher $\text{Overlap}_{\text{GT}}$ and lower $\text{Overlap}_{\text{Occ}}$, whereas textured occlusions (e.g., vegetation, smoke, rubble) exhibit lower consistency and greater attention leakage into occluded regions.

We report the saliency overlap metrics $\text{Overlap}_{\text{GT}}$ and $\text{Overlap}_{\text{Occ}}$ in Table 1. As expected, the saliency overlap metric reveal the following: gray occlusions preserve attention alignment with clean-image saliency maps better than textured occlusions. The classifier's focus under gray occlusion remains close to the object, showing higher $\text{Overlap}_{\text{GT}}$ and lower $\text{Overlap}_{\text{Occ}}$. Interestingly, smoke occlusions behave somewhat in between—being textured, but less structured—yielding moderate distraction compared to vegetation or rubble.

We report the saliency overlap metrics $\text{Overlap}_{\text{GT}}$ and $\text{Overlap}_{\text{Occ}}$ in Table 1. As expected, these results reveal clear differences in how occlusion type affects the classifier's attention. Under *gray occlusion*, the model maintains a strong alignment with its clean-image saliency, reflected in a higher $\text{Overlap}_{\text{GT}}$ and a lower $\text{Overlap}_{\text{Occ}}$. This indicates that the classifier's focus remains close to the visible object regions, with minimal attention shift into the occluded parts of the image.

In contrast, *textured occlusions* such as vegetation or rubble lead to a lower $\text{Overlap}_{\text{GT}}$ and a significantly higher $\text{Overlap}_{\text{Occ}}$, showing that the classifier's attention is more easily drawn into the occluded areas. Interestingly, *smoke occlusions* behave somewhat in between, being textured yet less visually detailed than vegetation or rubble, which results in only moderate distraction of the classifier's attention.

**Examples of saliency maps** To qualitatively demonstrate the attention of a classifier under different types of occlusion, we present examples of saliency maps on 5 occluded images in Figure 8. Note that for each image, the shape of the occlusion remains the same, only the type of the occlusion is changed. The first two columns show the clean, non-occluded image and the corresponding saliency map produced by the classifier. The remaining four columns depict the classifier's saliency under the four occlusion types. Empirically, and as expected, the saliency under *gray occlusion* (third column) remains concentrated on the still-visible parts of the object, with minimal shift into the occluded regions. In contrast, the last three columns reveal that the classifier increasingly struggles to maintain focus on the visible object areas when textured occlusions are introduced. This effect is particularly apparent for *vegetation* (fifth column) and *rubble* (seventh column) occlusions, where attention sometimes even drifts into the occluded regions.

### 6.2 Occlusion-agnostic segmentation

Having established how different occlusion types influence not only *how much* but also *where* the classifier attends under visual disturbance, we next turn to the question of where the occlusion actually lies. The goal is to localize occluded regions in an image without making any assumptions about what type of occluder is present. To this end, we employ **AnomalyDINO**, we use its generated anomaly maps to obtain occlusion segmentations. In doing so, we aim to determine whether occlusions can be detected in a fully occlusion-agnostic manner.

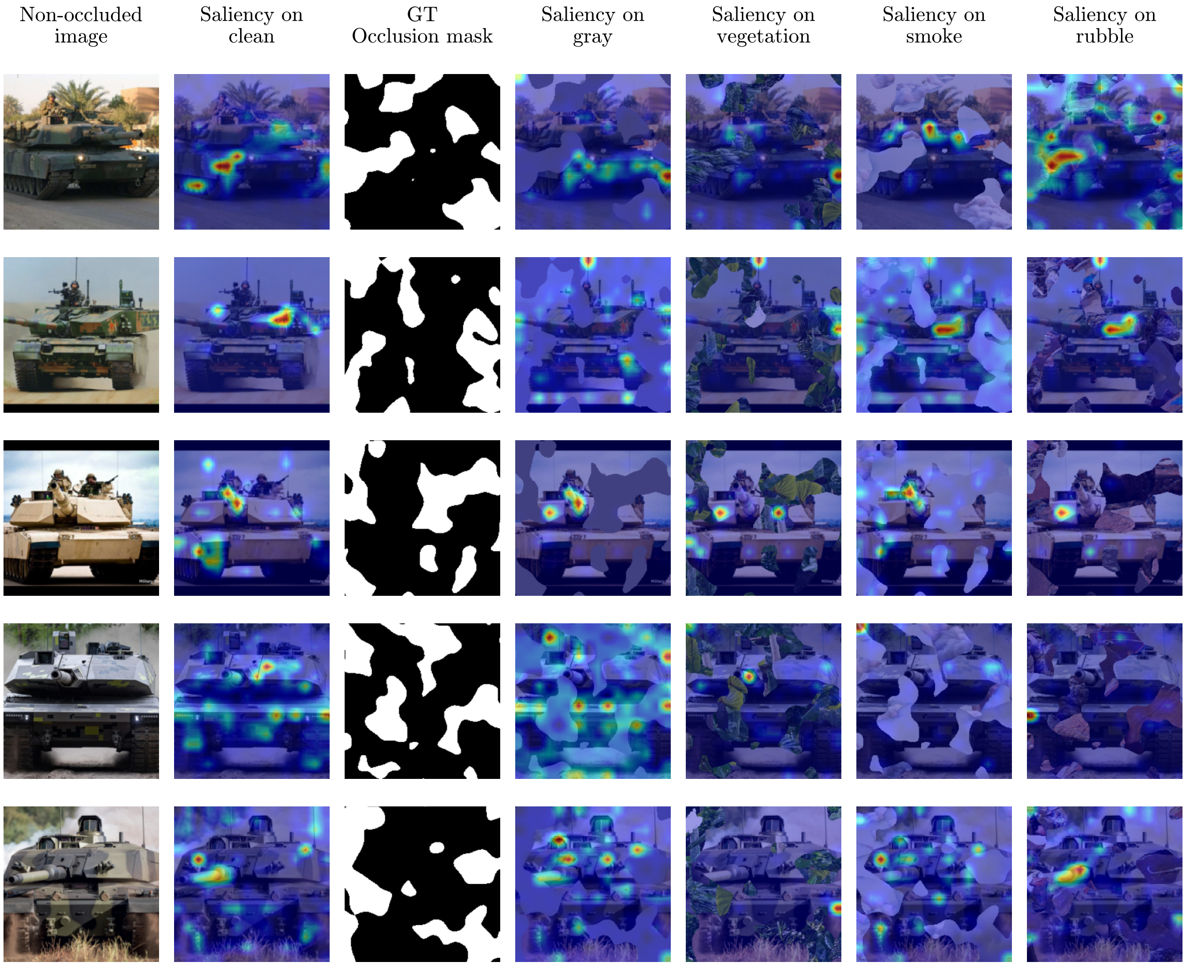


Fig. 8: **Qualitative visualization of saliency under different occlusion types.** The first column shows the original (unoccluded) images, and the second column displays their corresponding saliency maps. The occlusion mask applied to each row is shown in column 3 and remains the same across all occlusion types. Columns 4–7 present the occluded images: gray, vegetation, smoke, and rubble. Each is overlaid with its respective saliency map. Compared to gray occlusion, textured occlusions induce clear saliency shifts into the occluding regions, indicating that the model's attention is drawn toward the occluder rather than the object.

**Segmentation baseline** As a segmentation baseline, we use **OVSeg** [16], a recent open-vocabulary segmentation model that combines visual features with text embeddings from large-scale vision–language models. This enables OVSeg to produce segmentation masks for arbitrary textual prompts rather than being limited to a predefined label set. We exploit this flexibility to inject occlusion-specific information by prompting the model with occlusion-related terms (e.g., "vegetation", "smoke"). In doing so, OVSeg can identify regions corresponding to potential occluders without requiring any additional training. We employ OVSeg in a zero-shot setting using pretrained weights[4].

**From anomaly map to occlusion segmentation** Unlike conventional supervised segmentation methods, **AnomalyDINO** operates without explicit labels or occlusion-specific training. Instead, it detects outlier patches by comparing each patch's representation to a memory bank of non-anomalous patches. AnomalyDINO is inherently *occlusion-agnostic*: it is not trained to recognize any particular type of occluder but rather to identify irregularities that deviate from normal object appearance. Patches that diverge strongly from the memory bank receive high anomaly scores, which we interpret as indicators of occlusion. In essence, instead

[4] OVSeg weights: `ovseg_swinbase_vitL14_ft_mpt.pth`

of learning the visual characteristics of each occluder, AnomalyDINO highlights regions that do not conform to the expected object structure, thereby remaining robust across diverse occlusion types.

For AnomalyDINO, we build a memory bank of representative, non-occluded embeddings to serve as references during inference. For each of the 29 classes, we compute the class centroid in embedding space and populate the memory bank with the $k$ training images closest to this centroid. Experiments detailed in Appendix A show that performance saturates already at $k = 1$, meaning that a single reference image per class is sufficient. Hence, all reported results use $k = 1$.

| **Method** | **Vegetation** | | **Smoke** | | **Rubble** | |
|---|---|---|---|---|---|---|
| | mAUROC | mAP | mAUROC | mAP | mAUROC | mAP |
| **20% Occluded** | | | | | | |
| AnomalyDINO | **92.29±2.00** | **70.53±8.10** | **92.39±1.83** | **71.20±8.64** | **88.03±2.48** | **59.24±8.25** |
| OVSeg (prompt: "vegetation") | 79.29±13.38 | 65.46±15.04 | 44.10±5.81 | 20.58±1.08 | 50.80±4.90 | 25.47±9.44 |
| OVSeg (prompt: "smoke") | 45.51±3.99 | 20.96±1.29 | 59.54±10.37 | 34.67±15.48 | 48.88±6.27 | 23.01±5.18 |
| **40% Occluded** | | | | | | |
| AnomalyDINO | **88.73±4.10** | **79.27±8.15** | **88.72±4.17** | **81.27±6.21** | **85.09±5.26** | **74.92±7.77** |
| OVSeg (prompt: "vegetation") | 82.11±11.20 | 75.61±14.14 | 49.47±7.89 | 42.52±6.61 | 47.85±5.84 | 42.22±3.72 |
| OVSeg (prompt: "smoke") | 41.78±7.02 | 40.66±2.07 | 60.07±11.46 | 51.23±9.33 | 48.43±9.42 | 44.12±6.57 |
| **60% Occluded** | | | | | | |
| AnomalyDINO | **83.83±7.18** | **86.74±6.41** | **82.54±4.78** | **84.69±4.58** | **80.39±5.10** | **80.99±4.65** |
| OVSeg (prompt: "vegetation") | 73.24±12.35 | 77.27±10.36 | 49.07±11.53 | 62.29±6.82 | 49.39±6.48 | 61.78±3.59 |
| OVSeg (prompt: "smoke") | 42.94±5.48 | 59.34±1.76 | 63.23±8.29 | 67.42±5.86 | 44.34±8.12 | 60.31±3.57 |
| **80% Occluded** | | | | | | |
| AnomalyDINO | 60.16±10.10 | **90.05±4.40** | **66.68±9.49** | **88.57±3.92** | **67.42±10.84** | **87.32±4.12** |
| OVSeg (prompt: "vegetation") | **73.40±8.39** | 88.17±4.38 | 53.79±13.03 | 80.84±3.59 | 52.94±12.51 | 80.85±3.31 |
| OVSeg (prompt: "smoke") | 41.46±5.91 | 79.14±1.25 | 56.66±7.10 | 81.75±3.20 | 46.48±7.30 | 80.07±2.35 |

Table 2: **Mean AUROC (mAUROC) and mean Average Precision (mAP) across occlusion severities, for three segmentation methods.** Each column group corresponds to an occlusion type (vegetation, smoke, rubble), and each row group represents an occlusion level. For each metric–occlusion type pair, the **best result** is shown in bold. Scores where OVSeg is evaluated on the same occlusion type it was prompted with are underlined. Notably, *AnomalyDINO* achieves the highest segmentation performance across nearly all occlusion types and severities. While *OVSeg* performs reasonably well when prompted with "vegetation," it shows poor generalization to non-prompted occlusion types.

**Quantitative occlusion segmentation results** We evaluate occlusion segmentation under three types of occluders: vegetation, smoke, and rubble. We have the corresponding ground-truth occlusion mask, this allows for pixel-wise computation of the **Area Under the Receiver Operating Characteristic (AUROC)** and **Average Precision (AP)**. AnomalyDINO outputs pixel-level anomaly scores derived from patch distances, while OVSeg produces pixel-level probabilities. Both are used directly to compute AUROC and AP without applying any thresholding.

Table 2 reports the performance of segmenting the occlusion with, respectively: AnomalyDINO, OVSeg when prompted with "vegetation", and OVSeg when prompted with "smoke". For each method we present the mean AUROC (mAUROC) and mean Average Precision (mAP) achieved on image occluded with textured occlusion types *vegetation*, *smoke*, and *rubble*, for occlusion severities of 20%, 40%, 60%, and 80%. In both metrics, higher values indicate better segmentation performance. When comparing methods, **AnomalyDINO** consistently achieves the highest scores across all occlusion types and severities. As expected, it successfully segments occluded regions regardless of the specific occluder present, confirming its occlusion-agnostic design. Notably, AnomalyDINO's performance does drop (in terms of mAUROC) as occlusion severity increases.

In contrast, **OVSeg** performs well only when explicitly prompted with the correct occludesion type. For instance, when prompted with "vegetation", it segments vegetation occlusions reasonably well, but fails to generalize to other textures such as smoke and rubble. A similar pattern is observed when OVSeg is prompted with "smoke": performance is confined to that occlusion type, while other occlusion types remain poorly segmented. Remarkably, AnomalyDINO even surpasses OVSeg on the very occlusion types for which OVSeg was explicitly prompted, highlighting the robustness of its anomaly-based segmentation approach.

**Qualitative examples of occlusion segmentation** We present qualitative examples of occlusion segmentation for each method at 40% occlusion in Figure 9. The predicted segmentations are overlaid on the images using an opaque yellow mask. Each row corresponds to a different occlusion type (*vegetation*, *smoke*, and *rubble*), while the third, fourth and fifth columns represent: **AnomalyDINO**, **OVSeg** prompted with "vegetation", and **OVSeg** prompted with "smoke".

From these examples, it can be observed that OVSeg performs well only when prompted with the same occlusion type present in the image. For instance, when prompted with "vegetation", OVSeg successfully segments the vegetation occlusion (first row, fourth column 4), but fails to generalize to other occlusion types. In contrast, AnomalyDINO generalizes effectively across all occlusion types, producing consistent and accurate segmentations overall, though it slightly undersegments the *rubble* occlusion (bottom row). For simplicity, we applied a fixed threshold of $\tau = 0.5$ to the anomaly maps; some missed occluded regions may therefore be attributed to this sub-optimal threshold choice.

### 6.3 Masking occlusion segmentation with gray

Having established that occlusions can be segmented in an *occlusion-agnostic* manner using AnomalyDINO, we now turn to a more practical question: *how can this segmentation knowledge be leveraged to improve classification under occlusion?* In particular, we are interested in understanding whether removing the occlusion texture itself can benefit the classifier. Importantly, pixels within the segmented occluded regions are replaced with a neutral gray. By doing so, we effectively **suppress the distracting visual patterns** introduced by the occluder. We want to investigate whether reducing such visual distraction helps the classifier maintain better classification performance under increasing occlusion.

**Evaluating performance** To test this, we obtain occlusion segmentations by thresholding the continuous anomaly maps at $\tau = 0.5$. During inference, we preprocess the test images by replacing the segmented regions with a uniform gray. ensuring that the test samples resemble the gray-occluded images the model was trained on. The underlying intuition is that by converting any occlusion into a uniform, textureless region, the model is no longer tempted to overfit to the appearance of specific occluders, but instead learns to handle occlusions in a consistent, appearance-agnostic manner.

For evaluation, we compare two variants of the DINOv2-MLP classifier:

1. a **gray-trained** model, fine-tuned on gray-occluded images, where gray masking of the occlusion is performed; and
2. a **vegetation-trained** model, fine-tuned on vegetation-occluded images, without gray masking.

We expect the **gray-trained + gray masking** configuration to achieve the best performance across all occlusion types. This is partly because masking the occluded regions with gray reduces the visual distraction introduced by textured occlusions. Additionally, since this model was trained on gray-occluded images and evaluated on similarly gray-masked images (where the segmented occlusions are replaced with uniform gray), the occluding conditions during testing closely match those seen during training. In contrast, we expect the **vegetation-trained** model to perform well on *vegetation*-occluded images but to show reduced performance on *smoke* and *rubble* occlusions.

In all experiments, the occlusion severity level $p$ is assumed to be known, and both training and evaluation are performed at the same $p\%$ occlusion level.

Figure 10a presents the evaluation results for different occlusion types, shown from left to right for *vegetation*-, *smoke*-, and *rubble*-occluded images. Under all occlusion types, and across occlusion severities, the gray-trained model (gray) performs consistently. It achieves its highest accuracy on *smoke* occlusions, a texture that is visually similar to the neutral gray regions seen during training. The vegetation-trained model (green) shows limited cross-texture robustness, with modest gains on *smoke* and *rubble*, but still falls short of the gray-trained model's overall performance. Notably, fine-tuning on vegetation-occluded images (green) also achieves

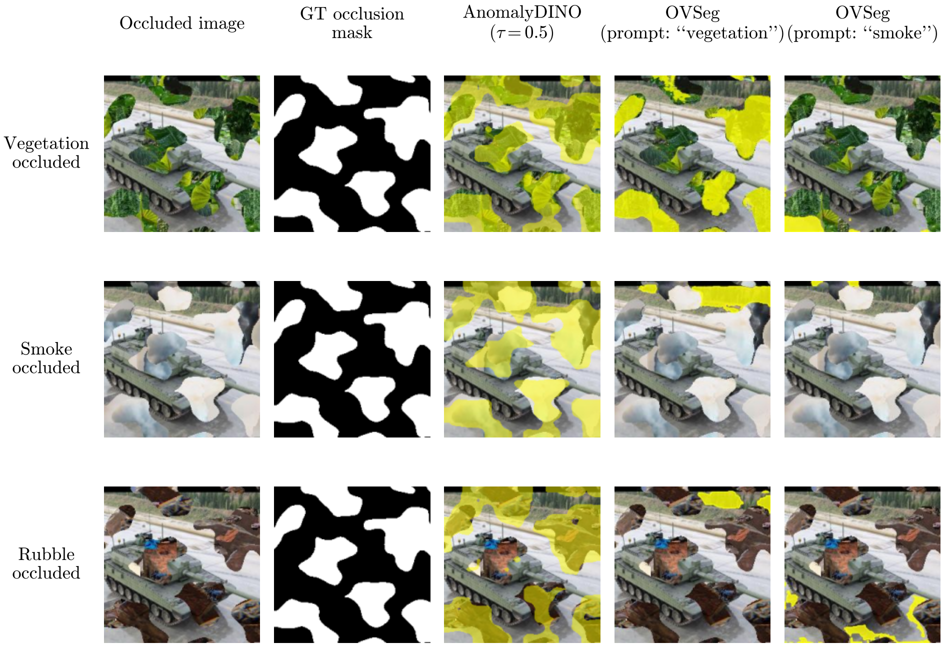
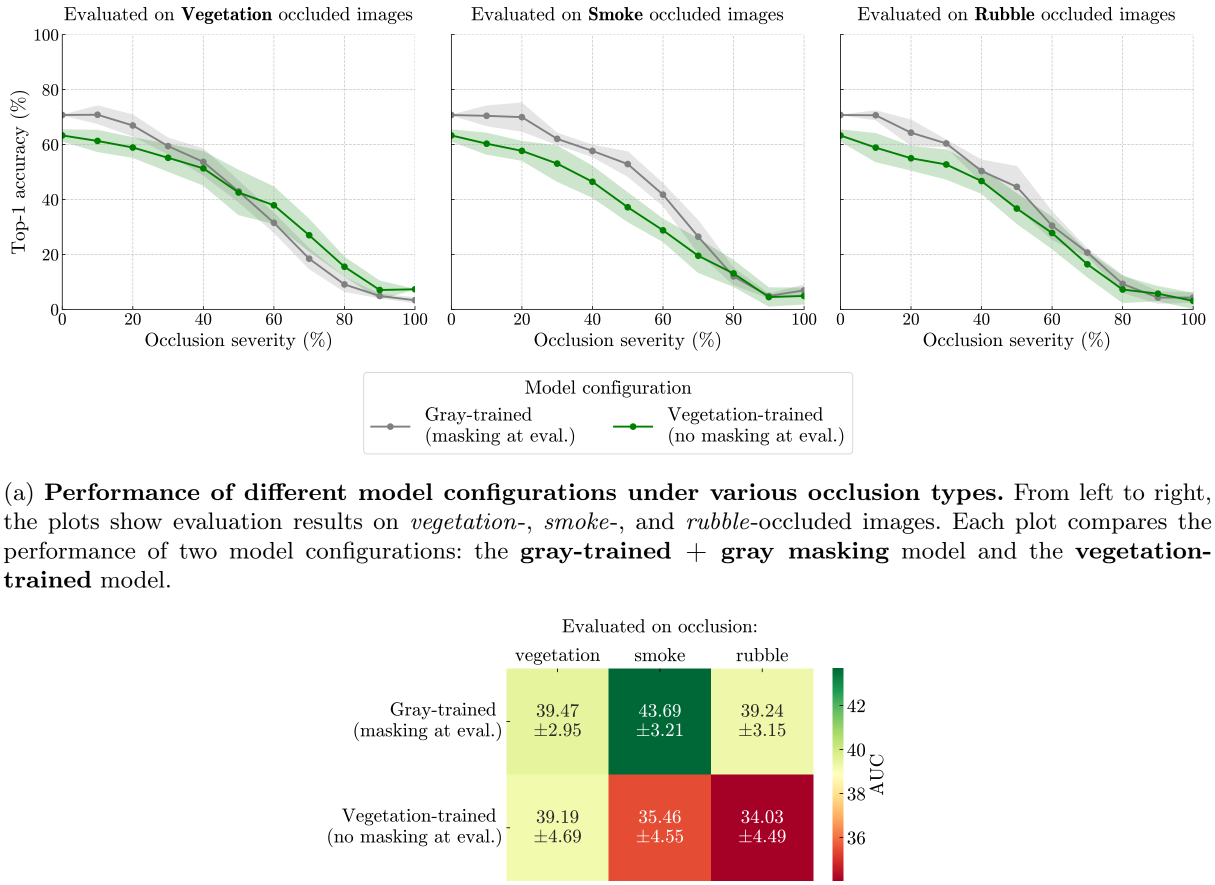


Fig. 9: **Qualitative comparison of occlusion segmentation at 40% occlusion severity.** Predicted segmentations are overlaid on the input images using opaque yellow. Each row corresponds to a different occlusion type (*vegetation*, *smoke*, and *rubble*), while the third, fourth, and fifth columns show results from **AnomalyDINO**, **OVSeg** prompted with "vegetation", and **OVSeg** prompted with "smoke", respectively. When prompted with "vegetation", **OVSeg** segments *vegetation* occlusion reasonably well but fails to generalize to other occlusion types. Segmenting *smoke* occlusion also appears challenging for **OVSeg**, even when it is prompted with "smoke". In contrast, **AnomalyDINO** produces consistent and accurate segmentations across all occlusion types, demonstrating strong generalization without occlusion-specific prompting.

sub-optimal performance on low- or non-occluded images. These results indicate that the **gray-trained + gray masking** configuration yields better and more generalizable performance across diverse occlusion types than fine-tuning on a single, texture-specific occlusion.



(a) **Performance of different model configurations under various occlusion types.** From left to right, the plots show evaluation results on *vegetation*-, *smoke*-, and *rubble*-occluded images. Each plot compares the performance of two model configurations: the **gray-trained + gray masking** model and the **vegetation-trained** model.

(b) **AUC$_{\text{occ}}$ scores across occlusion types and model configurations.** The heatmap presents the corresponding AUC$_{\text{occ}}$ values for the experiments shown in Figure 10a, with rows representing model configurations and columns denoting occlusion types (*vegetation*, *smoke*, *rubble*).

Fig. 10: **Evaluation of a gray-trained model (with gray masking applied at evaluation) and a vegetation-trained model (without gray masking) under different occlusion types.** The **gray-trained + gray masking** configuration achieves consistently higher AUC$_{\text{occ}}$ values across all occlusion types, while the **vegetation-trained** configuration attains comparable but slightly lower AUC$_{\text{occ}}$ on *vegetation* occlusions and performs notably worse on *smoke* and *rubble*.

Complementary results in Figure 10b summarizes AUC$_{\text{occ}}$ scores for each model–occlusion combination. The **gray-trained + gray masking** configuration achieves uniformly higher AUC$_{\text{occ}}$ values across all occlusion types, while the **vegetation-trained** configuration attains comparable but slightly lower AUC$_{\text{occ}}$ on *vegetation* occlusions and performs notably worse on *smoke* and *rubble*. These findings indicate that gray masking provides a more robust and generalizable strategy for handling diverse occlusion textures.

**Benefits of gray masking** As expected, masking occlusions with a neutral gray tone effectively suppresses the misleading visual cues introduced by textured occlusions. The robustness of the **gray-trained model** can, in part, be attributed to this setup: the model is trained on gray-occluded images and evaluated on similarly gray-masked images, where the segmented occluded regions are replaced with uniform gray. This consistency between training and evaluation conditions allows the model to generalize more effectively across diverse occlusion types.

The gray masking strategy hinges on a threshold $\tau$ applied to the continuous anomaly maps produced by AnomalyDINO. As of now, we stuck to a threshold of $\tau = 0.5$, but a different threshold value might be beneficial. This threshold $\tau$ controls how aggressively the occlusion is masked: a low $\tau$ yields extensive masking, while a high $\tau$ results in a more conservative approach. Here, we explore the delicate balance between masking enough to cover all occluded regions (maximizing true positives) and preserving unoccluded areas (minimizing false positives).

**Testing different levels of gray masking** To further understand the impact of gray masking, we investigate how the amount of masking influences classification performance. In particular, we explore the effects of masking either too much or too little of the occluded regions. To this end, we generate anomaly maps for images with textured occlusions ranging from 0% to 100% severity. Gray masking is then applied by thresholding the corresponding anomaly maps at fixed values $\tau \in \{0.1, 0.2, \ldots, 0.9\}$, where the regions exceeding the threshold are replaced with a neutral gray tone.

Each of the resulting masked images is evaluated using the same gray-trained DINOv2-MLP classifier as in the previous experiments, again under the assumption that the occlusion severity level is known. In addition to these fixed thresholds, we also compute a dynamic threshold $\tau_{\text{Otsu}}$ using **Otsu's method**, which adaptively determines $\tau$ based on the distribution of anomaly intensities within each image. This approach allows us to assess whether a data-driven threshold can automatically approximate an optimal masking level for improved classification under occlusion.

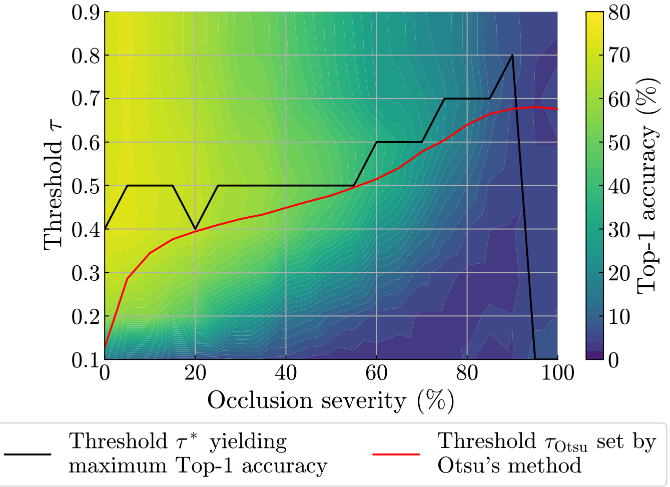


Fig. 11: **Effect of occlusion severity and masking threshold on classification performance.** Top-1 accuracy is shown as a heatmap, with yellow indicating higher accuracy and blue indicating lower accuracy. The vertical axis denotes the masking threshold $\tau$: lower values apply stronger gray masking, while higher values retain more of the original image. The **black line** indicates the optimal threshold $\tau^*$ for each occlusion severity, and the **red line** shows the automatically selected threshold $\tau_{\text{Otsu}}$. The optimal masking level varies with occlusion severity, with Otsu's method providing a close approximation.

Figure 11 visualizes how classification performance varies with occlusion severity and the degree of gray masking. The plot shows a heatmap of top-1 accuracy, where yellow indicates higher accuracy and blue indicates lower accuracy. The vertical axis represents the masking threshold $\tau$, where lower values correspond to stronger gray masking and higher values leave more of the original image visible. The **black line** denotes the optimal threshold $\tau^*$ that maximizes performance for each occlusion severity, while the **red line** indicates the threshold $\tau_{\text{Otsu}}$ automatically selected by Otsu's method. Together, these results highlight how the optimal level of masking shifts with occlusion severity.

Several trends stand out:

- **Overmasking** (low $\tau$) has little impact when occlusion severity is low but becomes increasingly detrimental as a larger portion of the image is occluded. Intuitively, at low occlusion levels, most of the object remains visible, so aggressive masking removes little information and does not significantly degrade performance.
- Consequently, **undermasking** (high $\tau$) is generally preferable to overmasking the occlusion, as it is better to leave some occluded regions unmasked than to wrongly mask visible parts of the object.
- The **optimal threshold** $\tau^*$ increases with occlusion severity, suggesting that as less of the object remains visible, the model benefits from more conservative masking that preserves the still visible regions. More conservative masking at higher occlusion severity helps preserve the limited visible parts of the object, leading to improved accuracy.
- The threshold $\tau_{\text{Otsu}}$, set by **Otsu's method**, tracks the optimal threshold remarkably well from roughly $20\%$ occlusion onwards, indicating that it can automatically identify a near-optimal masking level without manual tuning. At lower occlusion severities (below $20\%$), aggressive masking is less harmful, as most of the object remains visible and performance is relatively insensitive to the exact masking threshold.
- At $0\%$ occlusion, however, Otsu may overestimate the masked region (by setting a low $\tau_{\text{Otsu}}$), slightly hurting performance in the fully-visible case.

This analysis highlights that while the threshold $\tau$ is a sensitive hyperparameter, it can be effectively managed through adaptive selection. The fact that Otsu's method performs near-optimally across most occlusion severities is a useful property. This means gray masking can be applied in practice without requiring a per-severity tuning.

### 6.4 Fine-tuning for occlusion robustness

A common strategy to improve a model's robustness to occlusion is to fine-tune it on occluded images. However, this requires choosing a specific occlusion severity (or range of severities) to apply to the training data. Fine-tuning on occluded images is non-trivial, as a model optimized for one occlusion level may perform poorly at others. We expect this to occur because models adapt to the occlusion distribution they encounter during training. In other words, a model trained primarily on lightly occluded images may fail to generalize under heavy occlusion, while one trained on severely occluded inputs may underperform on unoccluded data.

**Evaluating fine-tuning strategies** We hypothesize that training on a dataset with low occlusion severity (i.e., where only a small portion of the image is masked) will lead to poor performance on highly occluded test images, and vice versa. To test this hypothesis, we fine-tune four instances of the DINOv2-MLP classifier on datasets with applied gray occlusions. Each model is trained under a distinct occlusion severity distribution and evaluated across all test-time occlusion levels ranging from $0\%$ to $100\%$. Images are occluded with gray, both during training and testing, to simulate idealized occlusion. This ensures removing visual information without introducing additional texture or color artifacts.

We consider four training configurations that differ in the range of occlusion severities used during fine-tuning. Each configuration is trained on either a fixed occlusion severity or a uniformly sampled range of severities. This setup allows us to systematically compare two key factors: (i) the influence of training on low versus high occlusion levels, and (ii) the effect of limited versus broad training exposure, in terms of their impact on classification performance across the full range of occlusion severities. The training configurations are as follows:

- **Narrow-range (0–20%)**: trained on images with occlusion levels uniformly sampled between 0% and 20%.
- **Wide-range (0–80%)**: trained on images with occlusion levels uniformly sampled between 0% and 80%.
- **Fixed (20%)**: trained on images with a constant 20% occlusion.
- **Fixed (80%)**: trained on images with a constant 80% occlusion.

**Effects of fine-tuning strategies** Figure 12 shows the performance of the four configurations, evaluated across all occlusion levels from $0\%$ to $100\%$. Models trained on low occlusion levels, *Narrow-range* (blue) and *Fixed 20%* (red) achieve high accuracy on clean or lightly occluded images but degrade rapidly as occlusion increases. Conversely, the *Fixed 80%* (orange) model performs best under severe occlusion but underperforms on low-occlusion and clean inputs, indicating that fine-tuning under heavily occluded images does not generalize to lower occluded images. The *Wide-range (0–80%)* (green) configuration maintains relatively stable accuracy across all severities, performing optimally under high occlusion and only slightly sub-optimally at low occlusion levels. This suggests that training across a broad range of occlusions yields greater robustness overall but sacrifices some peak accuracy under ideal (unoccluded) conditions.

Overall, models tend to specialize in the occlusion regime they are trained on. Low-occlusion trained models quickly lose accuracy as occlusion increases, while high-occlusion trained models fail to achieve satisfactory performance on non-occluded data. This confirms that fine-tuning for occlusion robustness is inherently non-trivial: we cannot simply train on a single occlusion severity and expect generalization across the full spectrum. The optimal robustness for a given occlusion level depends strongly on the severity distribution seen during training.

For completeness, a full matrix of fine-tuning and evaluation combinations, denoted $(p_{\text{test}}, p_{\text{train}})$, where $p_{\text{train}}$ is the training occlusion level and $p_{\text{test}}$ the test occlusion level—is included in Appendix C. This matrix confirms the same trend: performance peaks along the diagonal where training and test occlusion severities align.

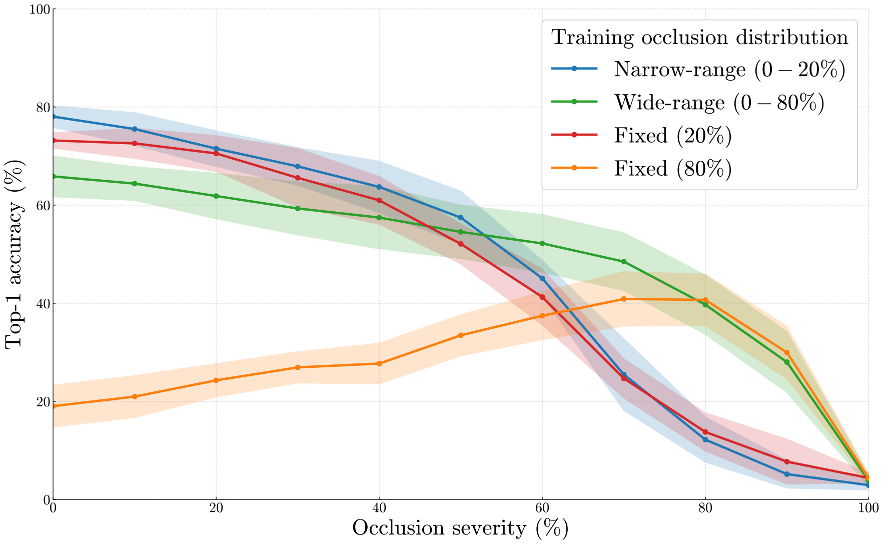


Fig. 12: **Performance of different fine-tuning configurations across occlusion levels.** Models trained on low occlusion levels, *Narrow-range* (blue) and *Fixed 20%* (red), perform well on clean or lightly occluded images but degrade as occlusion increases. The *Fixed 80%* (orange) model performs better under heavy occlusion but poorly on clean inputs, showing limited generalization. The *Wide-range (0–80%)* (green) configuration achieves the most balanced performance, maintaining stable accuracy across severities and demonstrating improved robustness at the cost of slightly reduced peak accuracy on unoccluded images.

**Severity-informed evaluation** This raises an important question: if the occlusion severity of a test image is known beforehand, could we improve performance by selecting the model that performs best under that severity?

Performance is generally highest when the fine-tuning occlusion severity matches the test severity, when $p_{\text{train}} = p_{\text{test}}$. In a few cases, particularly for range-trained models, slightly higher accuracy is achieved by selecting a neighboring model. For instance, when evaluating on 0% occlusion, the model trained on the $0$ – $20\%$ range slightly outperforms the model trained on clean data, indicating a mild regularization effect from a little exposure to occlusion. Since optimal performance typically lies close to the diagonal, we adopt $p_{\text{train}} = p_{\text{test}}$ as a practical simplification.

Figure 13 shows, for each test occlusion level $p$, the peak top-1 accuracy achieved by models fine-tuned on (left) a fixed $p\%$ occlusion and (right) a range of occlusions from $0 - p\%$. Both settings represent an oracle scenario in which the training occlusion distribution matches the test condition. Across all occlusion severities, models fine-tuned on fixed $p\%$ occlusions and those trained on the range $0 - p\%$ achieve nearly identical peak accuracy. However, the range-trained models exhibit more consistent performance across neighboring occlusion levels, reflected by a slightly higher mean accuracy and smaller standard deviation when averaged across models at each occlusion percentage. This indicates that training over a range of occlusions provides a mild regularization effect, leading to smoother and more stable generalization without compromising peak performance.

If the occlusion severity could be estimated at inference time, one could select the corresponding model fine-tuned for that level, a strategy we will refer to as severity-informed evaluation. In summary, models fine-tuned on a single occlusion severity become specialized to that specific level, whereas models trained across a range of occlusions (e.g., $0 - 80\%$) retain similar peak performance at high occlusion but generalize better to lower severities, resulting in more consistent robustness across occlusion conditions.

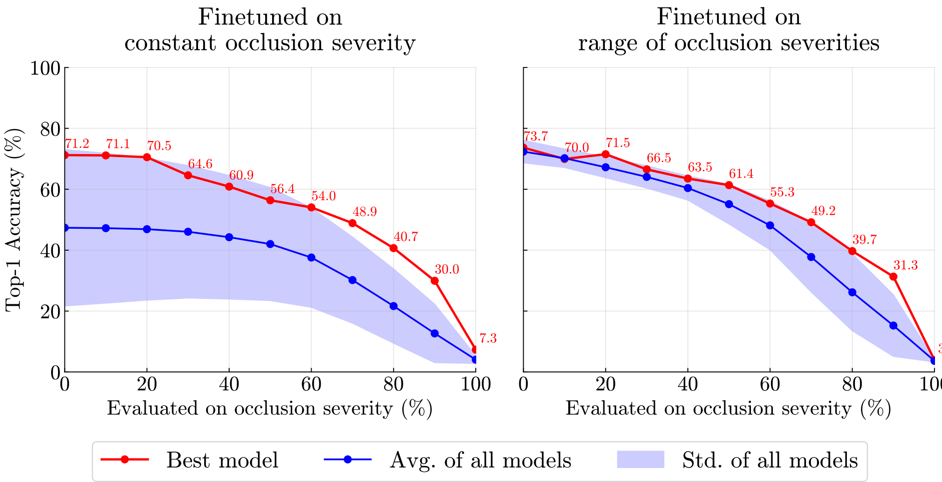


Fig. 13: **Effect of fine-tuning on different occlusion severity distributions.** For each test occlusion level $p$, the plots show the peak top-1 accuracy of models fine-tuned on (**left**) a fixed $p\%$ occlusion and (**right**) a range of occlusions from 0–$p\%$. Both settings represent oracle conditions where the training occlusion distribution matches the test scenario. Across all occlusion levels, models fine-tuned on fixed $p\%$ occlusions and those trained on the 0–$p\%$ range reach similar peak accuracy. Range-trained models, however, exhibit smoother performance across neighboring severities, with slightly higher mean accuracy and lower variance. This indicates that training over a range of occlusions acts as a mild regularizer, improving stability without sacrificing peak performance.

### 6.5 Severity-informed model selection

Previous experiments showed that no single model can be fine-tuned to perform optimally across all occlusion severities. Fine-tuning on images with a fixed occlusion level biases the model toward that specific condition, reducing its ability to generalize across different occlusion severities. Models trained on lightly occluded data perform poorly under heavy occlusion, while those fine-tuned on heavily occluded data struggle on clean or lightly occluded images.

Therefore, we explore whether model selection based on the estimated occlusion severity is a feasible alternative. Before such a strategy can be implemented, however, it is essential to determine whether occlusion severity can be reliably estimated from the anomaly maps produced by *AnomalyDINO*. These maps not only localize occluded regions but also encode pixel-wise anomaly intensities—a property that may be key to quantifying the extent of occlusion within an image.

If the magnitude of these anomaly values scales consistently with occlusion severity, then the anomaly maps could serve as a practical basis for estimating how severely an image is occluded, thereby enabling informed model selection at inference time.

**Estimating occlusion severity** We evaluated three strategies for estimating the occlusion severity $\hat{s}$ of test images affected by textured occlusions (*vegetation*, *smoke*, and *rubble*). For each image, we first apply the textured occlusion, generate the corresponding anomaly map using *AnomalyDINO*, and then use this map to estimate the occlusion severity. While many formulations for such an estimation are possible, we focus on the following three:

– $\hat{s}_{\text{mean}}$ (green), the mean anomaly score across the entire anomaly map;
– $\hat{s}_{\tau=0.6}$ (red), the proportion of pixels exceeding a fixed threshold $\tau$, here set to $0.6$;
– $\hat{s}_{\text{Otsu}}$ (blue), the proportion of pixels exceeding an adaptive threshold set by Otsu's method.

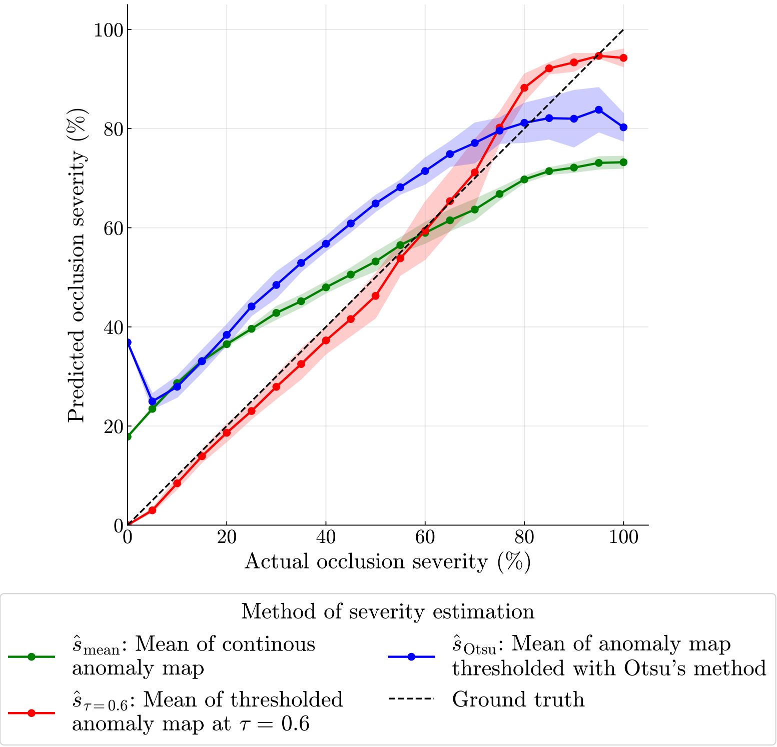


Fig. 14: **Estimated versus actual occlusion severity.** Estimated occlusion severity is shown for each ground-truth level. The mean anomaly score (green) overestimates severity at low occlusion and underestimates it at high occlusion. The proportion of pixels exceeding a threshold at $\tau = 0.6$ (red) yields accurate estimates that closely match the ground truth, while the Otsu-based estimator (blue) consistently overestimates severity.

Figure 14 reports the estimated occlusion severity $\hat{s}$ for each actual occlusion level. Taking the mean of the anomaly map (green) systematically *overestimates* severity at low occlusion and *underestimates* it at high occlusion. This suggests a nonlinear relationship between occlusion severity and mean anomaly score. As the occlusion severity increases, the mean anomaly score saturates and no longer scales proportionally, leading to an underestimation of the true occlusion severity. In contrast, thresholding the anomaly map at $\tau = 0.6$ (red) yields remarkably accurate estimates across all levels, aligning closely with the ground truth. Finally, the Otsu-based estimator (blue) consistently overestimates the true severity, as it is sensitive to local anomaly score fluctuations. This tends to select a threshold that classifies too much of the image as occluded.

**Selecting a model** Estimating occlusion severity directly from anomaly maps appears to be a feasible approach, particularly when using a simple fixed-threshold strategy (by thresholding at $\tau = 0.6$). However, an important question remains: does this estimated severity meaningfully help in selecting the most suitable model for a given image? Ultimately, the goal is not just to estimate occlusion accurately, but to improve classification performance under occluded conditions. An accurate severity estimate is therefore only valuable if it leads to better recognition results in practice.

In previous experiments, we observed that models trained on images with a fixed occlusion severity $p\%$ behave differently from models trained on a range of occlusion severities $0 - p\%$ when evaluated on out-of-distribution occlusion levels. To investigate this further, we compare two model selection strategies based on distinct training pools.

The first pool consists of models, each trained on images occluded with a fixed occlusion percentage $p\%$. The second pool contains models trained on images with a range of occlusion severities from $0\%$ up to $p\%$. the models in both pools are trained exclusively using gray occluded images.

For severity-informed model selection, we assign each test image to the model corresponding to its estimated occlusion severity $\hat{s}$. We then evaluate this choice by mapping $\hat{s}$ to the performance that the selected model achieved on the true occlusion severity $s$, as documented in Appendix C.

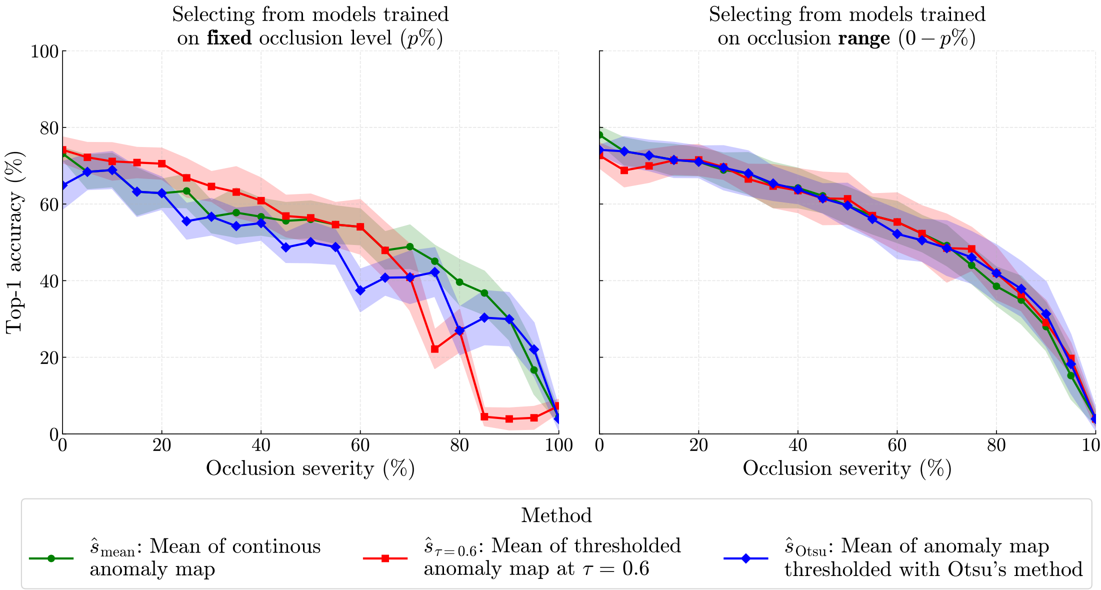


Fig. 15: **Classification performance achieved through severity-informed model selection.** The plots show the resulting classification performance when evaluating with models selected based on estimated occlusion severity, using three different estimation methods. (**Left**) When models are selected from a pool trained on fixed occlusion levels ($p\%$), performance remains sub-optimal. Even with near-perfect severity estimates (red), accuracy does not reach the upper bound, indicating that fixed-trained models are overly sensitive to small estimation errors and generalize poorly beyond their training severity. (**Right**) When selecting from a pool of range-trained models—each trained across a broader span of occlusion severities—performance is smoother and more consistent across estimation methods. Overall, training across a range of occlusions enhances generalization, mitigates the impact of estimation errors, and leads to more stable classification under varying occlusion conditions.

**Severity-informed performance** The performance of models selected using the three occlusion severity estimation methods is reported in Figure 15. When selecting from the pool of models trained on fixed occlusion severities ($p\%$; left), performance remains sub-optimal. As previous experiments have shown, these models perform best near their respective training occlusion level but degrade sharply when evaluated on out-of-distribution severities. Contrary to our expectations, even when using near-perfect severity estimates (red line), model selection fails to reach optimal performance. This suggests that models trained at fixed severities are overly sensitive to small inaccuracies in the estimated occlusion severity.

In contrast, selecting models from the $0-p\%$-trained pool (right) results in smoother and more consistent performance across all occlusion severities. This robustness is observed for all three estimation methods. Notably, approaches that slightly overestimate occlusion severity, such as the mean of the raw anomaly map (green) or the mean of the Otsu-thresholded anomaly map (blue), achieve performance nearly identical to that obtained with almost perfectly accurate severity estimates (red).

This finding aligns with earlier observations that training on a broader range of occlusion severities improves generalization to lower occlusion levels. Such broader training exposure likely compensates for minor estimation errors during severity-informed model selection, resulting in more stable and reliable performance.

**So which is best?** In summary, estimating occlusion severity proves beneficial—but only when combined with a sufficiently trained pool of models. Training on a range of occlusion severities ($0-p\%$) yields consistently reliable results, even when the severity estimates are not perfectly accurate. This indicates that robustness arises less from the precision of the severity estimation and more from the diversity of training exposure. In other words, slight imperfections in estimation are effectively absorbed by models trained across a broader severity spectrum.

Therefore, selecting models from the $0$–$p\%$-trained pool is the preferred strategy, as it delivers the most stable and consistent performance. Although estimating occlusion severity using the mean of the anomaly map tends to slightly overestimate the true severity, its simplicity and reliability make it a practical choice for our final experiment.

### 6.6 Evaluating OASIC

In previous experiments, we investigated two complementary strategies for improving robustness under occlusion: *gray masking* and *severity-informed model selection*. Gray masking mitigates the visual distraction of occluders and, when applied at inference time using the estimated occlusion map, ensures consistency between training and testing conditions. Meanwhile, estimating occlusion severity allows to select the model fine-tuned for the estimated occlusion level. Having demonstrated that both strategies improve performance individually, we now evaluate their combined effect within our complete method, *OASIC*.

**Evaluating performance across occlusion severities** Gray masking is applied according to the retrieved occlusion map, using Otsu's threshold $\tau_{\text{Otsu}}$, which was previously shown to provide a near-optimal separation between occluded and non-occluded regions. We therefore adopt this thresholding approach for the current experiment. For estimating occlusion severity, earlier results indicated that performance is robust to moderate estimation noise. Accordingly, we use $\hat{s}_{\text{mean}}$, which is the mean of the pixel-wise anomaly scores.

The following configurations are evaluated under progressively increasing occlusion severities:

– **Red**: performing **gray masking** combined with **severity-informed model selection** (i.e. our method, OASIC);
– **Blue**: performing **gray masking** with a **fixed model** trained on a broad occlusion severity range of $0-90\%$;
– **Green**: a **model trained on images containing vegetation occlusion**, where the training occlusion severity matches the evaluation severity $p\%$. We apply no gray masking;
– **Black**: a baseline **model trained exclusively on unoccluded images**.We apply no gray masking.

All models are tested across three textured occlusion types: vegetation, smoke, and rubble. Figure 16 visualizes their performances as occlusion severity increases. In the plot, the second, third, and fourth configurations serve as comparative baselines. The second configuration (blue), which applies gray masking but uses a single fixed model, illustrates the added benefit of severity-informed model selection. The third configuration (green), fine-tuned on vegetation-occluded images without gray masking, represents conventional training on a specific occlusion type. The fourth configuration (black), trained only on unoccluded images without any occlusion handling, provides a baseline to quantify the performance gains achieved by our occlusion-aware method.

**Results and observations** The results in Figure 16 reveal that our full method (red) consistently achieves the highest performance across all occlusion severities. This trend is further supported by the quantitative summary in Table 3, where using both gray masking and severity-informed model selection, attains the highest maximum accuracy and AUC$_{\text{occ}}$ among all evaluated methods. In terms of AUC$_{\text{occ}}$, our method improves $+6.36$ over the vegetatation-train configuration (green), and $+16.65$ over the clean-train configuration (black).

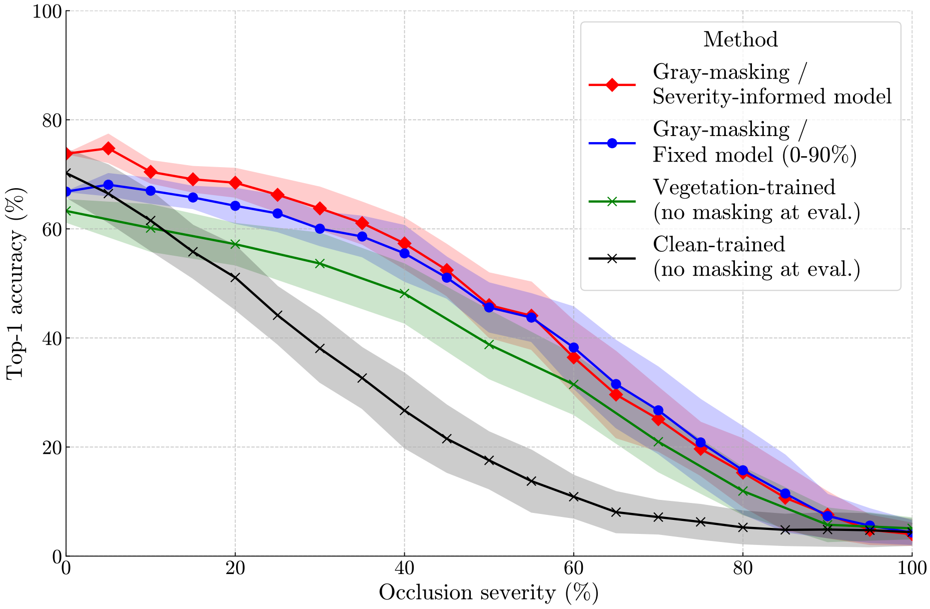


Fig. 16: **Occlusion robustness comparison.** Classification performance under increasing occlusion severity is shown for our method (red), which combines gray masking with severity-informed model selection, and three comparative configurations: gray masking with a fixed model (blue), a model trained on vegetation occlusions (green), and a model trained only on unoccluded images (black). Gray masking without severity-informed selection (blue) achieves comparable performance but performs slightly worse on lightly occluded images. Across all occlusion levels, our method (red) consistently attains the highest accuracy, demonstrating improved robustness to occlusion.

| **Method** | **Max. Accuracy (%)** | $\mathbf{AUC_{occ}}$ |
|---|---|---|
| Gray-masking / Severity-informed model (**red**) | $\mathbf{76.10}\pm\mathbf{0.12}$ | $\mathbf{42.59}\pm\mathbf{4.38}$ |
| Gray-masking / Fixed model (0-90%) (**blue**) | $68.10\pm2.03$ | $41.78\pm4.59$ |
| Vegetation-trained (no masking at eval.) (**green**) | $63.29\pm2.06$ | $36.23\pm4.57$ |
| Clean-trained (no masking at eval.) (**black**) | $70.24\pm4.41$ | $25.94\pm4.58$ |

Table 3: **Quantitative occlusion robustness comparison.** The table reports $\mathrm{AUC_{occ}}$ values and maximum accuracy for the four configurations shown in Figure 16. The **best results** for each metric are highlighted in bold. Our method (red) achieves the highest $\mathrm{AUC_{occ}}$ and maximum accuracy, demonstrating superior robustness across occlusion severities.

Together, these results confirm that integrating gray masking with severity-informed model selection provides a robust and reliable solution for handling variable occlusion levels.

Interestingly, the gray masking approach using a single model (blue) performs almost on par with the full method once occlusion exceeds 40%. At lower occlusion levels (0–40%), however, its performance lags slightly behind, indicating that severity-informed model selection is most beneficial when occlusion is moderate. This suggests that at high occlusion levels, the distinction between models trained for different severities becomes less relevant, likely because much of the visual information is already obscured.

Consistent with earlier experiments, the model trained on a specific occluder type (green) performs sub-optimally when tested across different occlusion types. Moreover, it shows reduced performance on images with little or no occlusion, highlighting the limited generalization of such specialized training.

**Concluding** In summary, combining gray masking with severity-informed model selection delivers the most consistent and robust performance under occlusion. When optimal accuracy at low occlusion levels is critical, it is beneficial to use the severity estimate to select the most suitable fine-tuned model. For applications requiring broader robustness, particularly when heavy occlusions are common, a simpler approach using gray masking with a single model trained across a wide severity range (e.g., $0$–$90\%$) is sufficient, though it comes with a slight reduction in performance on lightly occluded images.

## 7 Discussion

This study set out to improve fine-grained image classification performance under occlusion. To achieve this, we propose OASIC, a method that integrates occlusion-agnostic segmentation, gray masking, and severity-informed model selection. The approach leverages pixel-wise likelihoods of occlusion to identify and neutralize occluded regions by replacing them with a uniform gray mask, thereby suppressing distracting visual information that could mislead the classifier. In addition, we fine-tune a pool of classification models and dynamically select the most suitable one based on the estimated severity of occlusion.

### 7.1 Findings

**RQ1 In what ways does occlusion impact the reliability of image classification systems, considering both the reduction of observable object regions and the presence of distracting visual interference?**

Occlusion was found to substantially affect the reliability of image classification. Both gray and textured occlusions led to a reduction in classification accuracy; however, the impact of textured occlusion was notably greater. This indicates that the degradation in performance cannot be attributed solely to the reduction of visible object information. Instead, the visual characteristics of the occluding region itself appear to play a critical role.

An analysis of model saliency using EigenGradCAM further supports this observation. When exposed to textured occlusion, the classifier's attention was often diverted away from the remaining visible regions of the object and toward the occluded areas. This suggests that the presence of complex or patterned occluders introduces misleading visual cues that interfere with the model's ability to correctly interpret the scene.

In contrast, gray occlusion—being visually uniform and featureless—produced a smaller decline in accuracy. This form of occlusion limited the available object information but did not introduce competing or distracting features. As a result, the classifier maintained a more stable focus on the unobstructed parts of the object. These findings highlight that the reliability of image classification under occlusion is not only a matter of information loss but also of how visually distracting the occluding elements are.

**RQ2: To what extent can AnomalyDINO provide reliable occlusion segmentation and severity estimation?**

AnomalyDINO demonstrated strong and consistent performance in both occlusion segmentation and severity estimation tasks. Across all tested occlusion levels, it outperformed the baseline model (OVSeg, text-prompted with the occlusion type). While OVSeg, when prompted with "vegetation", showed reasonable segmentation performance on vegetation occluded images, it failed to generalize to other occlusion types, highlighting its dependence on prior knowledge of the occluder.

In contrast, AnomalyDINO operates in an occlusion-agnostic manner. It's approach is not to recognize particular occluders but instead to detect irregularities that deviate from the expected appearance of objects. This property enables it to identify occlusions of various kinds without explicit prior knowledge, providing a clear advantage in terms of generalization.

Furthermore, occlusion severity can be reliably estimated AnomalyDINO's anomaly maps, with near-perfect results obtained by measuring the proportion of pixels exceeding $\tau = 0.6$. This estimation enables the selection of severity-informed models, which outperform a single general model. Minor estimation errors were mitigated through a pool of models fine-tuned across broad occlusion ranges, maintaining robust performance despite slight severity estimation inaccuracies.

**RQ3 How can per-pixel occlusion likelihoods be integrated, through methods such as occlusion masking or severity-guided model selection, to improve fine-grained classification performance under occlusion?**

Combining gray masking with severity-informed model selection offers a robust and reliable solution for handling varying levels of occlusion. Quantitatively, our method improves $\text{AUC}_{\text{occ}}$ by $+6.36$ compared to training directly on occluded images and by $+16.65$ compared to fine-tuning on unoccluded images, demonstrating a substantial gain in occlusion robustness.

Applying gray masking suppresses the distracting visual patterns introduced by occluders, while severity-guided model selection enables to adapt to varying levels of information loss by selecting the most suitable model for a given occlusion level.

Distracting textures from occluders were successfully mitigated using a pixel-wise occlusion likelihood map derived from AnomalyDINO. Automatic adaptive thresholding of these maps using Otsu's method performed near-optimally, enabling gray masking to be applied without manual threshold tuning. This approach effectively reduced interference from complex occlusion patterns while preserving the visible object regions.

Estimating occlusion severity also proved beneficial, but only when combined with a sufficiently trained pool of models. Training across a range of occlusion severities (0–$p$%) produced consistent results, even when severity estimates were not perfectly accurate. This indicates that robustness arises less from the precision of the severity estimation itself and more from the diversity of exposure during training. In practice, small estimation inaccuracies were effectively absorbed by selecting from a pool of models fine-tuned across a broad severity spectrum.

### 7.2 Limitations

While the proposed anomaly-based segmentation approach proved effective, its performance remains dependent on the quality of the generated anomaly maps. In scenes with highly heterogeneous textures, the anomaly detection accuracy may degrade, potentially leading to less reliable occlusion segmentation. Moreover, severity estimation assumes a linear relationship between pixel-wise anomaly score and visibility loss, which may not always hold.

Another limitation concerns the granularity of severity estimation. The current approach estimates occlusion severity across the full image, whereas a more informative measure would focus specifically on the object region. Estimating the proportion of the object that is occluded, rather than the overall image area, would provide a more accurate and semantically meaningful severity assessment.

Furthermore, while AnomalyDINO offers strong generalization across unknown occluders, its advantages are most pronounced when the occlusion type is not known beforehand. In cases where the occluder is known, a segmentation model trained specifically on that occlusion type (e.g., vegetation or fog) may achieve comparable or even superior performance.

Finally, the optimal integration strategy depends on the target application. When high accuracy under low occlusion levels is critical, leveraging severity estimates to select a fine-tuned model yields the best results. However, for applications that prioritize robustness across a wide range of occlusions, a simpler approach—using gray masking with a single model trained across a broad severity range (e.g., 0–90%)—is sufficient, albeit with a modest reduction in performance for lightly occluded images.

### 7.3 Future work

The proposed framework offers potential for broader application beyond fine-grained recognition. Tasks where partial visibility is common, such as autonomous navigation, visual surveillance, or medical imaging under obstruction, could all benefit from anomaly-based occlusion estimation. Extending the current approach to these domains would further demonstrate its generality and robustness.

A promising direction for future research is the joint integration of occlusion estimation and classification within a single model architecture. Rather than treating occlusion reasoning as a separate preprocessing step, embedding it directly into the learning process could enable models to adaptively handle varying degrees of visibility. Similarly, extending the framework to temporal contexts, such as video data, could allow dynamic reasoning about occlusions that evolve over time.

While performance remains strong across most conditions, results indicate that heavy occlusions (around 60%) still pose a challenge. Addressing such cases may involve complementary techniques such as conformal prediction [24] or hierarchical classification [11], potentially in combination with the existing severity estimation. Integrating these uncertainty-aware methods could further improve the reliability of classification under extreme occlusion conditions.

**Acknowledgements** Please place your acknowledgments at the end of the paper, preceded by an unnumbered run-in heading (i.e. 3rd-level heading).

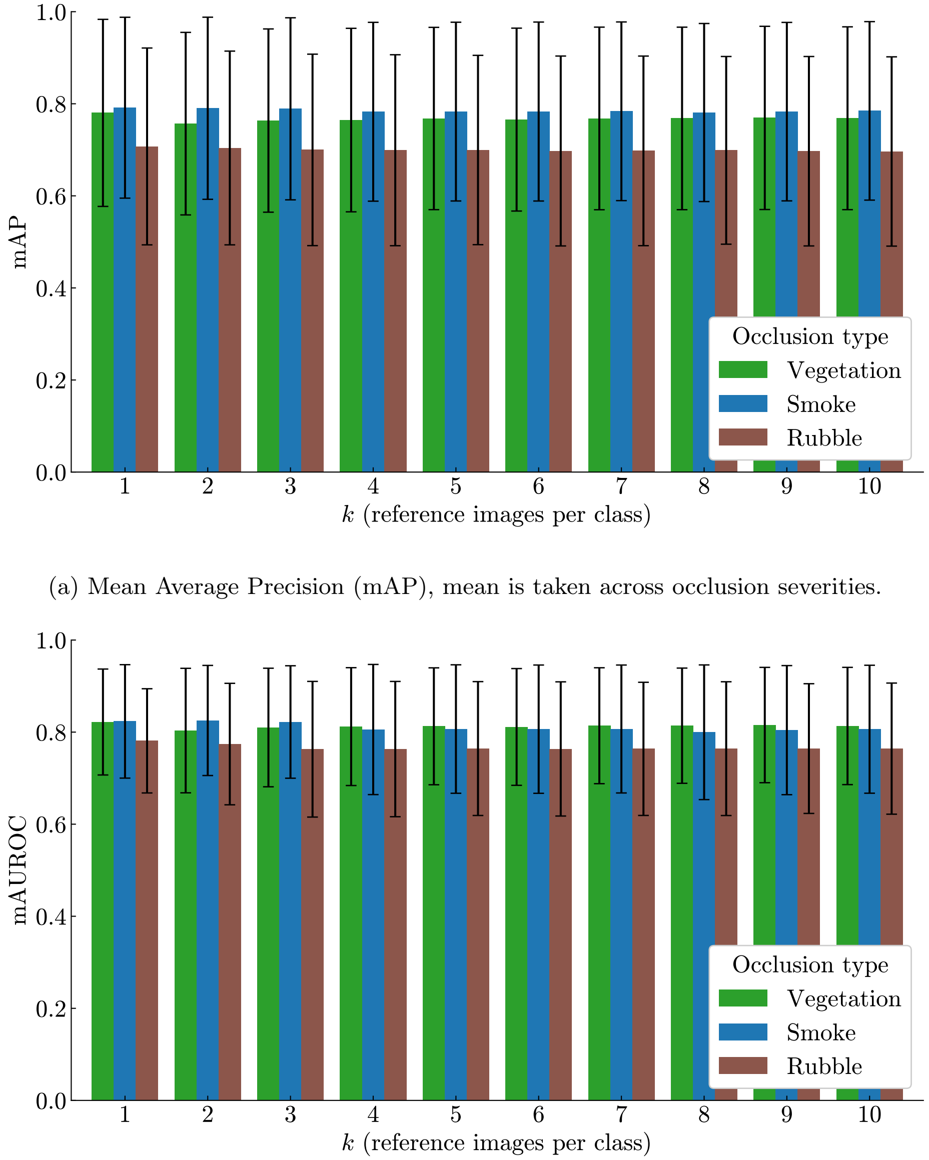
## References

1. Chen, J.N., Sun, S., He, J., Torr, P.H., Yuille, A., Bai, S.: Transmix: Attend to mix for vision transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12135–12144 (2022)
2. Damm, S., Laszkiewicz, M., Lederer, J., Fischer, A.: Anomalydino: Boosting patch-based few-shot anomaly detection with dinov2. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1319–1329. IEEE (2025)
3. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
4. 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)
5. Fawzi, A., Frossard, P.: Measuring the effect of nuisance variables on classifiers. In: BMVC. pp. 137–1 (2016)
6. French, R.M.: Catastrophic forgetting in connectionist networks. Trends in cognitive sciences **3**(4), 128–135 (1999)
7. Gildenblat, J., contributors: Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam (2021)
8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
9. Kassaw, K., Luzi, F., Collins, L.M., Malof, J.M.: Are deep learning models robust to partial object occlusion in visual recognition tasks? Pattern Recognition p. 112215 (2025)
10. 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. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
11. Koller, D., Sahami, M.: Hierarchically classifying documents using very few words. Tech. rep., Stanford InfoLab (1997)
12. Kong, X., Zhang, X.: Understanding masked image modeling via learning occlusion invariant feature. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6241–6251 (2023)
13. Kortylewski, A., He, J., Liu, Q., Yuille, A.L.: Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8940–8949 (2020)
14. Kortylewski, A., Liu, Q., Wang, H., Zhang, Z., Yuille, A.: Combining compositional models and deep networks for robust object classification under occlusion. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1333–1341 (2020)
15. Kumar Singh, K., Jae Lee, Y.: Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In: Proceedings of the IEEE international conference on computer vision. pp. 3524–3533 (2017)
16. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7061–7070 (2023)
17. Naseer, M.M., Ranasinghe, K., Khan, S.H., Hayat, M., Shahbaz Khan, F., Yang, M.H.: Intriguing properties of vision transformers. Advances in Neural Information Processing Systems **34**, 23296–23308 (2021)
18. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
19. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics **9**(1), 62–66 (1979)
20. Perlin, K.: An image synthesizer. ACM Siggraph Computer Graphics **19**(3), 287–296 (1985)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)

22. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
23. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2022)
24. Shafer, G., Vovk, V.: A tutorial on conformal prediction. Journal of Machine Learning Research **9**(3) (2008)
25. Shen, Y., Ding, H., Shao, X., Unberath, M.: Performance and nonadversarial robustness of the segment anything model 2 in surgical video segmentation. In: Medical Imaging 2025: Image-Guided Procedures, Robotic Interventions, and Modeling. vol. 13408, pp. 93–98. SPIE (2025)
26. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of big data **6**(1), 1–48 (2019)
27. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011. pp. 1521–1528. IEEE (2011)
28. Vashisht, A., Tekade, I., Shah, J., Sawarn, A., Yadav, D.S., Sontakke, P., Patil, R.: Effective segmentation of grape leaves using segment anything model 2. Smart Trends in Computing and Communications: Proceedings of SmartCom 2025, Volume 10 **10**, 375 (2025)
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
30. Wang, L., Cheng, S., Du, A., Wang, L., Zhang, L.: Occlusion simulation and token-constrained feature coupling network for occluded person re-identification. IEEE Internet of Things Journal (2025)
31. Xiao, M., Kortylewski, A., Wu, R., Qiao, S., Shen, W., Yuille, A.: Tdmpnet: Prototype network with recurrent top-down modulation for robust object classification under partial occlusion. In: European Conference on Computer Vision. pp. 447–463. Springer (2020)
32. Xu, K., Peng, Y., Zhou, J.: Uncover the body: Occluded person re-identification via masked image modeling. In: International Conference on Image and Graphics. pp. 241–253. Springer (2023)
33. 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)
34. Zavrtanik, V., Kristan, M., Skočaj, D.: Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8330–8339 (2021)
35. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
36. Zhang, Z., Xie, C., Wang, J., Xie, L., Yuille, A.L.: Deepvoting: A robust and explainable deep network for semantic part detection under partial occlusion. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1372–1380 (2018)
37. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenize. In: ICLR (2022)

## Supplementary Material

## A AnomalyDINO memory bank size



(a) Mean Average Precision (mAP), mean is taken across occlusion severities.

(b) Mean AUROC (mAUROC), mean is taken across occlusion severities.

Fig. 17: Occlusion segmentation performance of AnomalyDINO for different memory bank sizes ($k$ reference images per class). Aggregated, by taking the mean across occlusion percentages. For each class, the $k$ reference images are selected as those nearest to the class centroid in embedding space. Both metrics show that occlusion segmentation performance optimal with a single reference image per class ($k = 1$), with larger $k$ providing no further gains.

## B Saliency Overlaps

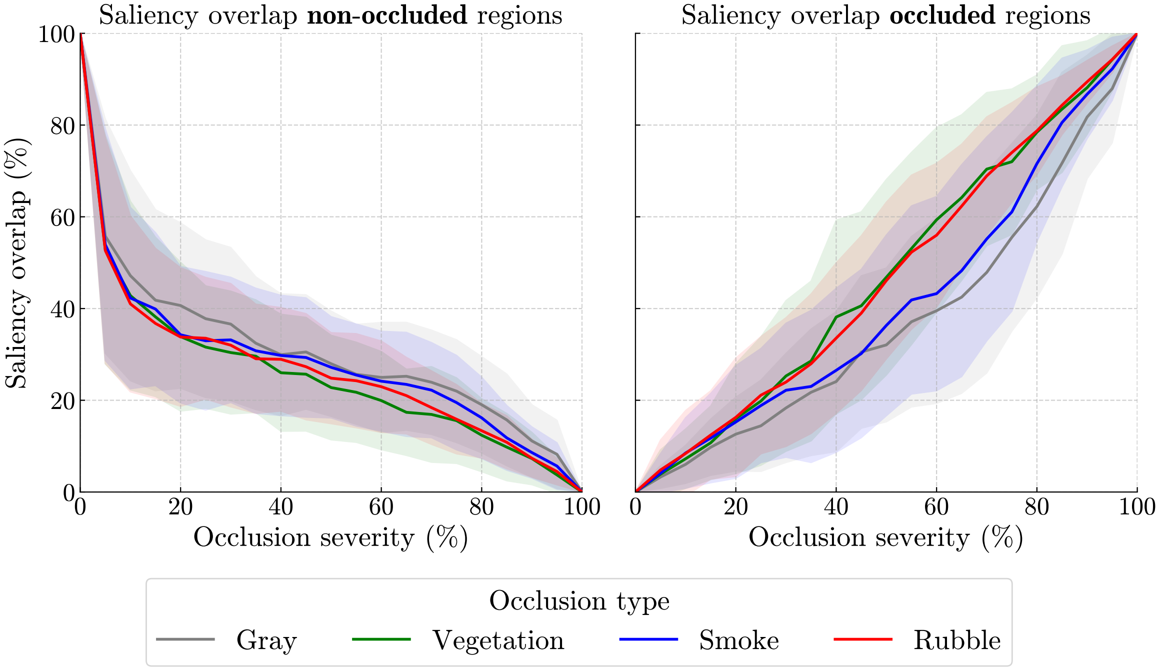
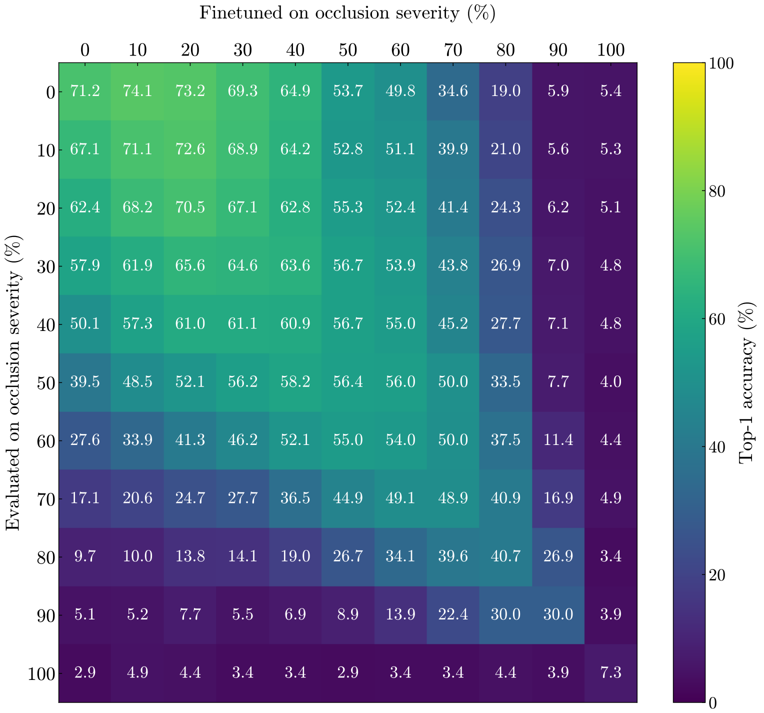


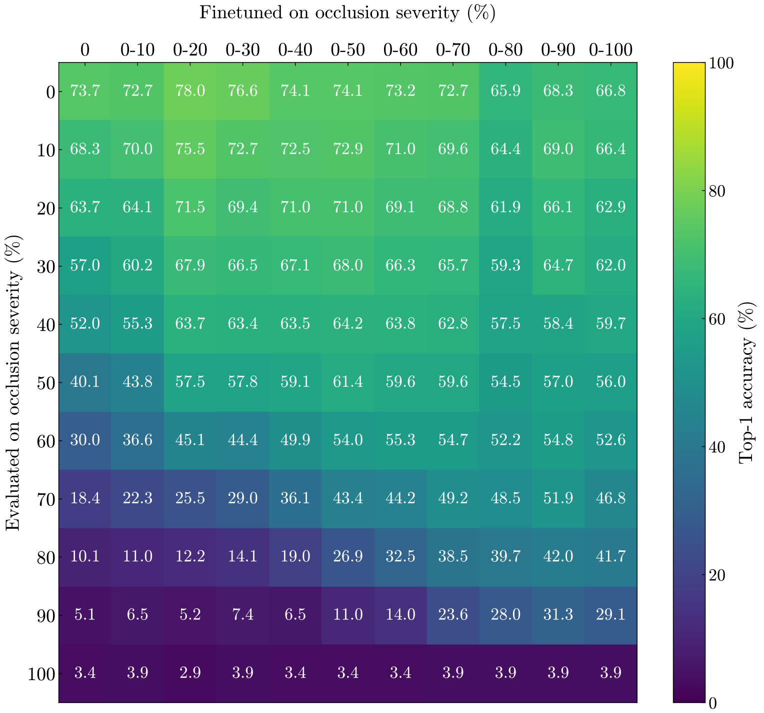
Fig. 18: Saliency overlap across occlusion severities for different occlusion textures (gray, vegetation, smoke, and rubble). Saliency maps were computed using Grad-CAM on a classifier trained on clean images. The left plot shows the overlap between the Grad-CAM saliency and the ground-truth saliency (computed on visible object regions), where a higher overlap indicates better localization of relevant features. The right plot shows the overlap between the Grad-CAM saliency and the occluded regions (ground-truth occlusion mask), where lower overlap is desirable. In both plots, the classifier achieves the most favorable overlap behavior for the gray occlusion condition.

## C Classification performance of fine-tuning on $p\%$ or $0 - p\%$ gray occlusion



| Evaluated \ Finetuned | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 71.2 | 74.1 | 73.2 | 69.3 | 64.9 | 53.7 | 49.8 | 34.6 | 19.0 | 5.9 | 5.4 |
| 10 | 67.1 | 71.1 | 72.6 | 68.9 | 64.2 | 52.8 | 51.1 | 39.9 | 21.0 | 5.6 | 5.3 |
| 20 | 62.4 | 68.2 | 70.5 | 67.1 | 62.8 | 55.3 | 52.4 | 41.4 | 24.3 | 6.2 | 5.1 |
| 30 | 57.9 | 61.9 | 65.6 | 64.6 | 63.6 | 56.7 | 53.9 | 43.8 | 26.9 | 7.0 | 4.8 |
| 40 | 50.1 | 57.3 | 61.0 | 61.1 | 60.9 | 56.7 | 55.0 | 45.2 | 27.7 | 7.1 | 4.8 |
| 50 | 39.5 | 48.5 | 52.1 | 56.2 | 58.2 | 56.4 | 56.0 | 50.0 | 33.5 | 7.7 | 4.0 |
| 60 | 27.6 | 33.9 | 41.3 | 46.2 | 52.1 | 55.0 | 54.0 | 50.0 | 37.5 | 11.4 | 4.4 |
| 70 | 17.1 | 20.6 | 24.7 | 27.7 | 36.5 | 44.9 | 49.1 | 48.9 | 40.9 | 16.9 | 4.9 |
| 80 | 9.7 | 10.0 | 13.8 | 14.1 | 19.0 | 26.7 | 34.1 | 39.6 | 40.7 | 26.9 | 3.4 |
| 90 | 5.1 | 5.2 | 7.7 | 5.5 | 6.9 | 8.9 | 13.9 | 22.4 | 30.0 | 30.0 | 3.9 |
| 100 | 2.9 | 4.9 | 4.4 | 3.4 | 3.4 | 2.9 | 3.4 | 3.4 | 4.4 | 3.9 | 7.3 |

(a)



| Evaluated \ Finetuned | 0 | 0-10 | 0-20 | 0-30 | 0-40 | 0-50 | 0-60 | 0-70 | 0-80 | 0-90 | 0-100 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 73.7 | 72.7 | 78.0 | 76.6 | 74.1 | 74.1 | 73.2 | 72.7 | 65.9 | 68.3 | 66.8 |
| 10 | 68.3 | 70.0 | 75.5 | 72.7 | 72.5 | 72.9 | 71.0 | 69.6 | 64.4 | 69.0 | 66.4 |
| 20 | 63.7 | 64.1 | 71.5 | 69.4 | 71.0 | 71.0 | 69.1 | 68.8 | 61.9 | 66.1 | 62.9 |
| 30 | 57.0 | 60.2 | 67.9 | 66.5 | 67.1 | 68.0 | 66.3 | 65.7 | 59.3 | 64.7 | 62.0 |
| 40 | 52.0 | 55.3 | 63.7 | 63.4 | 63.5 | 64.2 | 63.8 | 62.8 | 57.5 | 58.4 | 59.7 |
| 50 | 40.1 | 43.8 | 57.5 | 57.8 | 59.1 | 61.4 | 59.6 | 59.6 | 54.5 | 57.0 | 56.0 |
| 60 | 30.0 | 36.6 | 45.1 | 44.4 | 49.9 | 54.0 | 55.3 | 54.7 | 52.2 | 54.8 | 52.6 |
| 70 | 18.4 | 22.3 | 25.5 | 29.0 | 36.1 | 43.4 | 44.2 | 49.2 | 48.5 | 51.9 | 46.8 |
| 80 | 10.1 | 11.0 | 12.2 | 14.1 | 19.0 | 26.9 | 32.5 | 38.5 | 39.7 | 42.0 | 41.7 |
| 90 | 5.1 | 6.5 | 5.2 | 7.4 | 6.5 | 11.0 | 14.0 | 23.6 | 28.0 | 31.3 | 29.1 |
| 100 | 3.4 | 3.9 | 2.9 | 3.9 | 3.4 | 3.4 | 3.4 | 3.9 | 3.9 | 3.9 | 3.9 |

(b)

Fig. 19: Visualization of model robustness across occlusion severities under gray occlusion. (a) shows results for models trained at a fixed occlusion level ($p\%$) and evaluated across increasing occlusion severities. (b) shows results for models trained on progressively larger occlusion ranges ($(0 - p\%)$) evaluated across increasing occlusion severities. This comparison highlights the effect of range-based training on generalization to unseen occlusion levels.