# EXPLORING ALEATORIC UNCERTAINTY IN OBJECT DETECTION VIA VISION FOUNDATION MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Datasets collected from the open world unavoidably suffer from various forms of randomness or noiseness, leading to the ubiquity of aleatoric (data) uncertainty. Quantifying such uncertainty is particularly pivotal for object detection, where images contain multi-scale objects with occlusion, obscureness, and even noisy annotations, in contrast to images with centric and similar-scale objects in classification. This paper suggests modeling and exploiting the uncertainty inherent in object detection data with vision foundation models and develops a data-centric reliable training paradigm. Technically, we propose to estimate the data uncertainty of each object instance based on the feature space of vision foundation models, which are trained on ultra-large-scale datasets and able to exhibit universal data representation. In particular, we assume a mixture-of-Gaussian structure of the object features and devise Mahalanobis distance-based measures to quantify the data uncertainty. Furthermore, we suggest two curial and practical usages of the estimated uncertainty: 1) for defining uncertainty-aware sample filter to abandon noisy and redundant instances to avoid over-fitting, and 2) for defining sample adaptive regularizer to balance easy/hard samples for adaptive training. The estimated aleatoric uncertainty serves as an extra level of annotations of the dataset, so it can be utilized in a plug-and-play manner with any model. Extensive empirical studies verify the effectiveness of the proposed aleatoric uncertainty measure on various advanced detection models and challenging benchmarks.

## 1 INTRODUCTION

Deep learning has witnessed remarkable success in a wide range of scenarios and applications for predictive performance, such as image classification Liu et al. (2021); Dosovitskiy et al. (2021); Tolstikhin et al. (2021); He et al. (2016), semantic segmentation Xie et al. (2021); Strudel et al. (2021), and object detection Carion et al. (2020); Zhang et al. (2022); Zhu et al. (2021a); Ren et al. (2015); He et al. (2017). Datasets collected from the open world unavoidably suffer from various randomness or noiseness Kendall & Gal (2017); Cui et al. (2022), resulting in ubiquitous uncertainty inherent in the data (i.e., *aleatoric* uncertainty or data uncertainty Der Kiureghian & Ditlevsen (2009); Hüllermeier & Waegeman (2021)). Quantifying such uncertainty is pivotal for comprehending the inherent fluctuations within the training data, which enables the construction of more resilient models that can accommodate and flexibly respond to conditions characterized by inherent uncertainty.

Compared to images with centric and similar-scale objects in classification benchmarks, images in object detection datasets are typically scene-centric and contain multiple objects in varying scales. Especially, some objects are accompanied by occlusion, obscureness, and even noisy annotations due to limited resources and time in the data collection process Liu et al. (2022) (as observed in Fig. 1). Naturally, the aleatoric uncertainty arises in object detection tasks. However, the majority of aleatoric uncertainty quantification methods target classification or regression problems Kendall & Gal (2017); Chang et al. (2020); Depeweg et al. (2018); Zhang et al. (2024a), with few focusing on the fundamental and challenging object detection. To bridge the gap, we aim to investigate aleatoric uncertainty at the *detection level*, i.e., in the context of object detection.

It is almost impossible for human annotators to compare samples within the dataset and quantify each sample's aleatoric uncertainty due to unaffordable time and resource costs. When discriminative features from object instances are salient and obvious, we consider the aleatoric uncertainty to be low, as such instances can be easily detected and assigned to their semantic classes. However, when

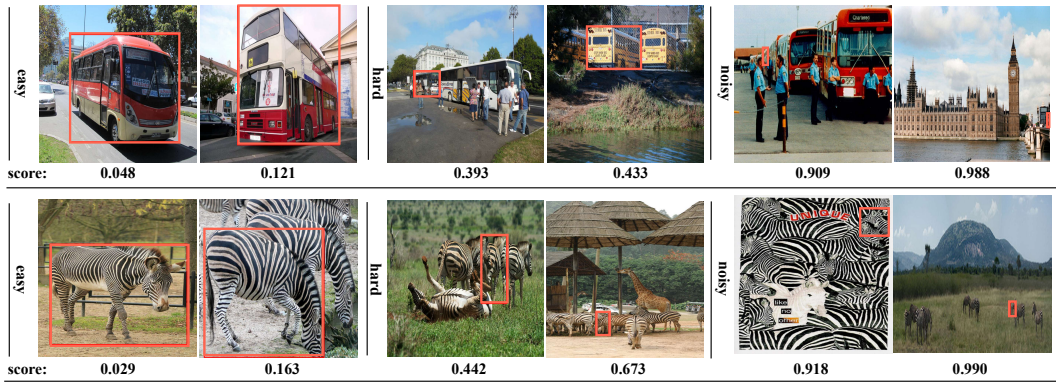



Figure 1: Visualization of scoring objects with corresponding uncertainty scores in training images of MS-COCO. The orange bounding box is the annotated ground truth. "Hard" objects suffer from occlusion or obscureness within an image, and "noisy" ones have misleading bounding boxes.

some of these features are occluded or missing, aleatoric uncertainty increases, making it more challenging to localize and classify such instances. Vision foundation models have learned rich and well-structured features from large-scale training data, enabling them to compare samples from diverse perspectives. In this paper, we opt for SAM Kirillov et al. (2023) as the foundation model and bridge the gap in utilizing SAM to characterize aleatoric uncertainty at the detection level. SAM was trained on the expansive SA-1B dataset Kirillov et al. (2023) that contains more than 1 billion masks spread over 11 million carefully curated images and has established superior performance in addressing open-world vision tasks. Unlike vision foundation models CLIP Radford et al. (2021) and DINOv2 Oquab et al. (2023), SAM received direct supervision in solving dense prediction tasks. Its vision encoder outputs high-resolution feature maps, which is beneficial for processing object detection datasets where objects can vary significantly in size.

In light of this, we proceed to capture the feature of each object instance in the feature space of SAM for measuring aleatoric uncertainty. Building on top of the existing ground-truth bounding boxes and class labels, we perform bounding box-based feature pooling to get a feature vector per object. While SAM was trained in a class-agnostic manner, semantically similar instances are found to be closely crowded together in its feature space. In recent work Xiaoke et al. (2024), one can directly assign semantics and generate captions based on SAM's output embeddings via some text feature mixture and decoder. Based on these observations, we employ a class-conditional Gaussian distribution to model the feature distribution and derive a Mahalanobis distance-based uncertainty score as a measure of aleatoric uncertainty. As shown in Fig. 1, the proposed uncertainty score effectively captures pertinent data characteristics such as object difficulty and noise level, aligning well with human cognitive level.

Furthermore, we devise two practical and crucial usages related to aleatoric uncertainty: uncertainty-aware sample filtering and loss regularization. We can utilize them as proxy tasks to examine the quality of aleatoric uncertainty and enhance detection performance. Firstly, we introduce a quantile function based on aleatoric uncertainty scores to abandon noisy samples that may mislead model training, as well as redundant samples within sub-populations grouped by uncertainty scores to improve training efficiency. Secondly, we propose a sample adaptive training objective that incorporates uncertainty-aware entropy to regularize the binary cross-entropy loss, which can balance easy and hard samples more knowledgeably compared to typical focal loss Lin et al. (2017) and entropy regularization Pereyra et al. (2017).

Aleatoric uncertainty measure can serve as additional annotations of training data thus it can be employed for any model in a plug-and-play way. We conduct extensive empirical studies on challenging benchmarks: MS-COCO Lin et al. (2014) and BDD100K Yu et al. (2020), corresponding to natural and self-driving scenarios, respectively. These studies were performed using various advanced detectors, e.g., CNN-based YOLOX Ge et al. (2021) and FCOS Tian et al. (2019), and transformer-based Deformable DETR Zhu et al. (2021a) and DINO Zhang et al. (2022), to verify the effectiveness of the aleatoric uncertainty measure. We first show that the sample adaptive regularizer incorporated data uncertainty can improve detection performance regarding averaged precision and recall. Furthermore, significant performance gains are observed when aleatoric uncertainty is exploited to abandon noisy samples, and our uncertainty-aware filter strategy outperforms uniform

sampling for redundant instances filtering. Finally, we conduct informative ablation studies to show the robustness of hyperparameters and further explore the potential of aleatoric uncertainty.

## 2 Related Works

**Wide applications of SAM.** SAM Kirillov et al. (2023) is a vision foundation model designed to address dense prediction tasks by outputting instance masks and parts within regions of interest specified via visual prompts such as points and bounding boxes. Its strong generalization capabilities across domains have enabled a wide spectrum of downstream use cases. While SAM itself only provides class-agnostic masks, it can be utilized after semantic-aware object detection to generate masks for each bounding box. For instance, Grounded-SAM Ren et al. (2024) that connects SAM with Grounding DINO Liu et al. (2023b) is a strong open-world object detection and segmentation model with text prompts. Exploiting caption models Li et al. (2022; 2023a) or image tagging models Huang et al. (2023a); Zhang et al. (2023); Huang et al. (2023b) to get semantic descriptions of images and further use them as text prompts, Grounded-SAM serves as an effective auto-labeling tool. Additionally, SAM has been employed in segmentation tasks within industrial defect segmentation Cao et al. (2023); Li et al. (2024) and medical image segmentation Zhang et al. (2024b).

In recent work Xiaoke et al. (2024), SAM was found to know semantics implicitly. Instead of starting from a semantic-aware object detection model, SAM can do captioning and assign semantic classes to the generated masks through a combination of text feature mixture and a text decoder following its vision encoder. We target a novel use case of SAM: annotating the aleatoric uncertainty of each training sample, which is distinct from the annotations of usual bounding boxes and masks. We benefit from the implicit semantic knowledge in the feature space of SAM's vision encoder. Nevertheless, our use case does not rely on an extra text decoder or feature mixture.

**Feature space density modeling.** Understanding data distribution provides insights into data structure, the generation of additional samples following the same distribution, and out-of-distribution (OOD) detection. Leveraging feature extractors trained to provide compact and informative data representations, feature space density modeling has been proven more effective for tasks like OOD detection, e.g., Kirichenko et al. (2020); Ren et al. (2021); Liang et al. (2022). Based on the familiarity hypothesis in Dietterich & Guyer (2022), relying on rich features is particularly beneficial. Due to their large-scale training sets, the vision encoders of foundation models like SAM effectively fulfill this purpose. While various density modeling techniques have been developed and OOD detection is one of the main use cases, we introduce a new use case: aleatoric uncertainty estimation, which is distinct from OOD detection. Although we employ a standard density modeling method, the achieved gains highlight the potential in this novel application.

**Aleatoric uncertainty.** In deep learning, uncertainty can be classified into two categories: *aleatoric* or data uncertainty and *epistemic* or model uncertainty Der Kiureghian & Ditlevsen (2009); Kendall & Gal (2017); Hüllermeier & Waegeman (2021). Depeweg et al. (2018) propose a decomposition method of uncertainty to capture aleatoric uncertainty from the predictive distribution of Bayesian neural networks with latent input variables. Similarly, Kendall & Gal (2017) developed a technique using MC-dropout Gal & Ghahramani (2016) to independently characterize both uncertainty components. Zhang et al. (2024a) propose a prediction-model-agnostic denoising approach to estimate aleatoric uncertainty for regression by augmenting a variance approximation module under the assumption of the zero mean distribution of data noise. Chang et al. (2020) introduces a data uncertainty-aware method for face recognition by learning feature (mean) and uncertainty (variance) simultaneously in the feature embedding. Prior works mainly estimate aleatoric uncertainty for classification or regression tasks by predictive uncertainty decomposition on task-specific data distribution and training model. This explores the ability of vision foundation models trained on diverse data to be used to quantify data uncertainty from a data distribution perspective.

## 3 Aleatoric Uncertainty Quantification in Object Detection

As data collection and annotation processes inevitably suffer from varying degrees of corruption, aleatoric uncertainty (i.e., data uncertainty) is ubiquitous in real-world datasets. Accurately characterizing data uncertainty can help us better understand training data to utilize it more efficiently and reliably, especially for modern large-scale datasets. To quantify data uncertainty, we leverage SAM to extract the feature of each object instance and model the training data distribution by fitting a multivariate Gaussian distribution in the feature space. Prior work Cui et al. (2024) has shown the effectiveness of Gaussian distribution modeling on classification tasks. We anticipate that easy

samples with low uncertainty will be closely crowded together, while hard/noisy ones with high uncertainty will be far away from the population and more dispersed. A similar intuition is utilized to quantify uncertainty in the classification literature in previous works Van Amersfoort et al. (2020); Mukhoti et al. (2023). From the perspective of density estimation within feature distribution, we derive a Mahalanobis distance-based uncertainty score to represent aleatoric uncertainty. We detail the whole process in Algorithm 1.

**Multivariate Gaussian distribution.** The training dataset consists of the image-label pairs: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ with $x_i \in \mathbb{R}^d$ and $y_i = \{b_j, c_j, s_j\}_{j=1}^{M}$, and $y_i$ represents the set of ground-truths for each image where $b_j \in \mathbb{R}^4$ and $s_j$ is the bounding box and binary mask for each object instance $z_j$, and $c_j \in \{1, \ldots, K\}$ is the corresponding class. Let $V(\cdot)$ denote the feature map layer of the vision encoder in SAM, and we can employ it to obtain each image's feature embedding $V(x_i)$. Building on $V(x_i)$ and corresponding ground-truths: $b_j$ or $s_j$, we further acquire each object's feature vector: $V(z_j)$. The conditional Gaussian distribution with the class $k$ can be defined as:

$$P(V(z) \mid c = k) = \mathcal{N}\left(V(z) \mid \mu_k, \Sigma\right), \tag{1}$$

where $\mu_k$ is the mean vector for class $k$, and $\Sigma$ is an averaged covariance matrix shared by all classes for all training samples. Specifically, we can empirically estimate them by

$$\begin{aligned} \mu_k &= \frac{1}{N_k} \sum_{j:c_j=k} V\left(z_j\right), \\ \Sigma &= \frac{1}{N} \sum_{k} \sum_{j:c_j=k} (V(z_j) - \mu_k)(V(z_j) - \mu_k)^\top, \end{aligned} \tag{2}$$

where $N_k$ is the number of training samples (i.e., object instances) with the label $c_j = k$.

**Mahalanobis distance-based uncertainty score.** Leveraging the class-conditional Gaussian distributions fitted above, we measure the Mahalanobis distance between training object $z$ and the corresponding class-conditional Gaussian distribution to represent the aleatoric uncertainty of each object in the training set., i.e.,

$$\mathcal{M}(z_j|c_j) = -\left(V(z_j) - \mu_{c_j}\right)^\top \Sigma^{-1} \left(V(z_j) - \mu_{c_j}\right). \tag{3}$$

The Mahalanobis distance $\mathcal{M}(z_j|c_j)$ measures the distance between an object and the centroid of the category $c_j$. A small $\mathcal{M}(z_j|c_j)$ indicates that the object has typical features of the subpopulation belonging to this class and boils down to low data uncertainty. Oppositely, the object with the high $\mathcal{M}(z_j|c_j)$ tends to contain ambiguous information (i.e., insufficient identifying characteristic) or noisy annotation (i.e., ambiguous bounding box or even wrong class label). In order to more conveniently exploit data uncertainty, we employ a scaling procedure to transform the Mahalanobis distance to a range of $(0, 1)$, which is achieved through a combination of log transformation and min-max normalization techniques:

$$d(z_j|c_j) = \frac{\log\left(\mathcal{M}(z_j|c_j)\right) - \min\limits_{j:c_j=k}\{\log \mathcal{M}(z_j|c_j)\}}{\max\limits_{j:c_j=k}\{\log \mathcal{M}(z_j|c_j)\} - \min\limits_{j:c_j=k}\{\log \mathcal{M}(z_j|c_j)\}}, \tag{4}$$

where the Mahalanobis distance belonging to each class is individually normalized to $(0, 1)$. Fig. 2 illustrates the distribution of $d(z_j|c_j)$ for the training data of MS-COCO Lin et al. (2014), and we also show the distribution by categories in Appendix. It is evident that a small percentage (approximately 5%-10%) of samples exhibit high uncertainty scores, implying the presence of noisy objects within the dataset. Additionally, a significant proportion of objects in the MS-COCO dataset are characterized as difficult/hard, as evidenced by the high density of uncertainty scores within the range of $0.5 - 0.6$.

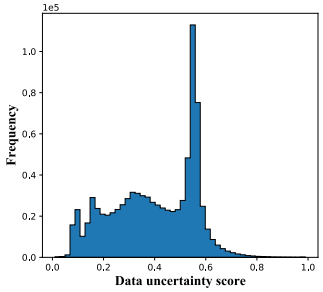



Figure 2: The histogram of $d(z_j|c_j)$ for MS-COCO.

Furthermore, we give some sorting examples and their uncertainty scores belonging to classes "bus" and "zebra" in Fig 1, see Appendix for more visual examples. We can observe a high level of agreement between human visual perception and MD-based data uncertainty scores. In conclusion, our empirical investigation suggests that:

- The low data uncertainty represents an easy sample that can be readily recognized by humans or models due to abundant and unbroken features.

- The objects with medium uncertainty scores are often located in distant positions or partially obscured within an image, posing challenges for accurate classification and detection.
- The objects with high uncertainty often indicate low-quality samples, which may stem from unrecognizable instances or misleading annotated bounding boxes and categories. These instances are prone to being regarded as data noise due to their ambiguity or inconsistency.

**Algorithm 1:** Aleatoric uncertainty quantification for object detection

**Input:** Training dataset $\mathcal{D}$, the feature map layer of vision encoder in SAM: $V(\cdot)$

```
for x_i, y_i in D do
    Get feature embedding V(x_i) for each x_i;
    for b_j, s_j, c_j in ground-truths set y_i do
        Compute the feature vector V(z_j) of each object based on ground-truths b_j or s_j;
        Add the feature vector V(z_j) to the feature set V;
    end
end
Compute the mean vector and averaged covariance matrix using Eqn. (2);
Compute Mahalanobis distance of each object using Eqn. (3);
Obtain the final uncertainty score of each object by Eqn. (4) and save all uncertainty scores.
```

## 4 A Recipe for Aleatoric Uncertainty in Object Detection

This section explores the practical usage of aleatoric uncertainty in object detection. Based on estimated aleatoric uncertainty, we propose various data filtering strategies, which aim to remove underlying noisy and redundant objects from the training dataset, leading to more efficient and reliable model training. To further enhance the predictive performance, we develop an uncertainty-aware regularizer and incorporate it into the loss function. Moreover, these two usages can also serve as a proxy for examining the quality of estimated aleatoric uncertainty. It is worth noting that we can calculate the per-object uncertainty score beforehand and treat it as an offline proxy, so the proposed uncertainty-aware usages do not take any additional computational overhead during the model training. In particular, uncertainty scores can serve as an extra level of annotations of the training set and be utilized for any model in a plug-and-play way.

### 4.1 Aleatoric uncertainty-aware data filtering

**Filtering out noisy objects.** As shown in Fig. 1 and 2, some objects have incomplete discriminative features or incorrect annotations, which can damage model training and lead to poor predictive performance. Given this, we propose discarding possible noisy samples that are harmful to model learning. Specifically, we employ a quantile function to discard objects with high uncertainty scores during model training. Let $F$ denote the cumulative distribution function (CDF) of uncertainty scores over all classes, and then we can use the inverse function of CDF $F^{-1} : [0, 1] \rightarrow d(z_j|c_j)$ to represent its quantile function:

$$F^{-1}(p) = \inf \{d : p \leq F(d(z_j|c_j))\}. \tag{5}$$

After that, we retain objects $\mathcal{D}^*$ that are smaller than the specific quantile $p$ (e.g, $p = 95\%$) used for model training, i.e.,

$$\mathcal{D}^* = \left\{ z_j | d_j \leq F^{-1}(p) \right\}_{j=1}^{N*M}. \tag{6}$$

Considering the class-imbalanced issues Lin et al. (2014) in the MS-COCO dataset, we also try discarding noisy objects according to per class, i.e., first calculating the inverse function of CDF of each class $c_j$, referred to as $F^{-1}_{j:c_j=k}(p)$, and then retaining the objects $\mathcal{D}^*$ that meet:

$$\mathcal{D}^* = \left\{ \left\{ z_j | d_j \leq F^{-1}_{j:c_j=k}(p) \right\}_{j:c_j=k}^{N_k} \right\}_{k=1}^{K}, \tag{7}$$

where $N_k$ is the number of object instances with the label $c_j = k$.

**Filtering out redundant objects.** Object detection datasets, such as MS-COCO, typically contain numerous similar objects with common patterns. Thus, an additional useful application of the uncertainty score is eliminating potentially redundant objects from the training set. Objects with closely clustered uncertainty scores within each class often exhibit similar or common patterns. Consequently, the model may only need to learn from a subset of these instances to achieve satisfactory

performance. In this spirit, we can select a certain proportion of objects, known as valuable samples, from each sub-population with close uncertainty scores to enhance training efficiency.

Concretely, we group the uncertainty score of each object into $M$ interval bins for each class (each of size $1/M$) and randomly throw away $p\%$ objects in each bin. We use 10 bins in this work, and we provide the results of more bins in Table A1 in the Appendix. Let $B_m^{c_j}$ be the set of indices of samples with class $c_j$ whose uncertainty score falls into the interval $I_m = \left(\frac{m-1}{M}, \frac{m}{M}\right]$, and the object set that we retain: $\mathcal{D}^*$ can be denoted as:

$$\mathcal{D}^* = \left\{ \left\{ z_j | j \in B_m^{c_j=k} \right\}_{m=1}^{M} \right\}_{k=1}^{K}, \tag{8}$$

### 4.2 Aleatoric uncertainty-aware Regularization

The uncertainty score serves as a valuable tool for characterizing each object's difficulty and noise level, as demonstrated in Fig. 1. Therefore, it is worth exploring how to leverage this knowledge to enhance model performance. The object detection models usually optimize multiple losses, e.g., $L = L_{\text{cls}} + L_{\text{box}} + L_{\text{obj}}$, and the standard training loss formulation is data uncertainty agnostic. The previous work, such as focal loss Lin et al. (2017), primarily focuses on fitting hard samples and mitigating overfitting to easy samples. It is defined as $L_{\text{FL}} = -(1 - P_t)^{\gamma} \log(P_t)$, where $P_t$ is the model's predictive probability of the ground-truth class and $\gamma$ is a predefined coefficient designed to alleviate the model overfitting to the already confident (i.e., $P_t$ close to 1) majority class. Yet, the focal loss is sensitive to coefficient $\gamma$ and may lead to inappropriate or even harmful regularization for some samples based on the predicted probability.

To address this issue, we incorporate data uncertainty score $d(z_j|c_j)$ into classification loss $L_{\text{cls}}$ and propose an uncertainty-aware entropy to regularize the binary cross-entropy loss. Besides, prior work Mukhoti et al. (2020) has demonstrated that cross-entropy loss equipped with a maximum-entropy regularizer can be interpreted as the lower bound of focal loss, resulting in the ability of the proposed uncertainty-aware entropy regularizer to ensure the optimal performance of the model. As a result, we arrive at the sample adaptive classification loss:

$$\mathcal{L}_{cls} = -\frac{1}{N * M} \sum_{j=1}^{N*M} \left( (1 - c_j) \log(1 - f_\theta(z_j)) + c_j \log(f_\theta(z_j)) - \beta d(z_j|c_j) \mathcal{H}[f_\theta(z_j)] \right), \tag{9}$$

where $f_\theta(z_j)$ and $\mathcal{H}[f_\theta(z_j)]$ refer to the predictive binary probability distribution and corresponding entropy for the object $z_j$. $\beta$ is a predefined coefficient to control the strength of entropy regularization, which generally ranges from $(0.1, 0.3)$. Since we can estimate the per-object uncertainty score once before training, the proposed training objective scarcely introduces additional computing overhead.

## 5 Experiments

To verify the effectiveness of the proposed aleatoric uncertainty measure in conveying valuable information about the dataset, we report the predictive performance when employing it for both data filtering and sample adaptive regularization. We mainly present primary experimental results on the challenging benchmark MS-COCO Lin et al. (2014) for the bounding box detection task and use Deformable DETR Zhu et al. (2021b) and anchor-free YOLOX Ge et al. (2021) as the object detection models. In the Appendix, we provide more results for other detection models, such as FCOS Tian et al. (2019) and DINO Zhang et al. (2022). We also validate our method on the challenging self-driving dataset: BDD100K Yu et al. (2020) and other VFMs. Moreover, we ablate the robustness of the proposed uncertainty-aware entropic regularizer to hyper-parameters.

**Datasets.** The 118k train set (train2017) and the 5k validation set (val2017) of COCO 2017 are utilized for training and evaluating the model on the bounding box (bbox) detection task. COCO 2017 comprises 80 classes and encompasses a diverse range of scenes, including indoor and outdoor environments, urban and rural settings, as well as various lighting and weather conditions. The training set contains, on average, 7 instances per image, with a maximum of 63 instances observed in a single image. These instances span a wide range of sizes, from small to large.

**Metrics.** To evaluate the prediction quality, we report averaged precision (AP) and recall (AR) over IoU thresholds, $\text{AP}_{50}$, $\text{AP}_{75}$, and $\text{AP}_L$, $\text{AP}_M$, $\text{AP}_S$ for various-scale objects.

**Implementation Details.** In our study, we mainly employ two typical detection models as detectors: the transformer-based Deformable DETR (trained up to Epoch 50, with a 4-scale setup) and the CNN-based anchor-free YOLOX (specifically, YOLOX-S and YOLOX-M versions), and model details are summarized in Table 1. We also report the performance of more detectors in the Appendix. Deformable DETR surpasses previous DETR Carion et al. (2020) in both performance and efficiency, achieving better performance than DETR (especially on small objects) with 10×less training epochs by combining the best of the sparse spatial sampling of deformable convolution. YOLOX transforms the traditional YOLO detector, such as YOLOv3 Redmon & Farhadi (2018), into an anchor-free method and enhances it with a decoupled head and the proposed label assignment strategy SimOTA, thereby achieving state-of-the-art performance. As for implementation details, e.g., data preprocessing, experimental settings, etc., we completely follow the original paper. Moreover, we do not use strong data augmentation techniques such as Mixup Zhang et al. (2018) for all experiments. For the hyper-parameters in the proposed training loss 9, we set $\beta$ as 0.2 and 0.3 for YOLOX and Deformable DETR, respectively.

Table 1: Model details.

| Model | Params | GFLOPS |
|---|---|---|
| YOLOX-S | 9M | 26.8 |
| YOLOX-M | 25M | 73.8 |
| D-DETR | 40M | 265 |

### 5.1 Performance on uncertainty-aware regularizer

Table 2 demonstrates the performance comparison between binary cross-entropy with a constant weighting (Entropy) and uncertainty-aware entropy (UA-entropy) for YOLOX-S, YOLOX-M, and Deformable DETR. The proposed uncertainty-aware entropic regularizer is obviously the top-performing one and leads to a consistent improvement across all detection models. Notably, the performance gain is also prominent for the small-scale models, i.e., YOLOX-S and YOLOX-M, indicating that the proposed data uncertainty measure can convey valuable information about the dataset to model learning. More importantly, the superior performance gain of the proposed sample adaptive regularizer on small-scale models holds significant implications for real-world model deployment. Conversely, regularizing each sample with equal entropy shows only slight improvement or even deteriorates model performance, especially for the small-scale detector YOLOX-S (-0.88% AP). Moreover, the proposed method also achieves significant performance gain on more advanced Deformable DETR with focal loss, implying that the proposed training objective effectively combines data uncertainty to more reasonably balance the learning of difficult and easy samples. We also report the performance of other detectors (i.e., FCOS and DINO) in the Appendix, showing the consistent performance gain.

Table 2: Performance comparison of uncertainty-aware entropy (UA-entropy) and constant entropy regularizer (Entropy) on COCO valset.

| Model | Method | AR | AP | $AP_{50}$ | $AP_{75}$ | $AP_L$ | $AP_M$ | $AP_S$ |
|---|---|---|---|---|---|---|---|---|
| YOLOX-S | Vanilla | 53.92 | 39.43 | 57.62 | 42.53 | 52.53 | 43.24 | 21.22 |
| | Entropy | 52.97 | 38.55 | 55.83 | 41.58 | 51.71 | 42.43 | 20.34 |
| | UA-entropy | **54.26** | **39.85** | **58.66** | **43.13** | **52.84** | **43.67** | **22.05** |
| YOLOX-M | Vanilla | 57.92 | 44.34 | 62.27 | 47.98 | 58.32 | 48.33 | 26.69 |
| | Entropy | 58.22 | 44.41 | 62.54 | 48.22 | 58.31 | 48.71 | 26.84 |
| | UA-entropy | **58.86** | **45.33** | **63.78** | **49.12** | **58.99** | **49.94** | **27.97** |
| Deformable DETR | Vanilla | 67.44 | 46.22 | 65.23 | 50.00 | 61.73 | 49.21 | 28.82 |
| | Entropy | 67.23 | 46.10 | 65.01 | 49.25 | 61.06 | 48.34 | 28.17 |
| | UA-entropy | **68.43** | **47.59** | **66.84** | **51.96** | **62.57** | **50.66** | **30.34** |

### 5.2 Performance on uncertainty-aware data filter

**Filtering of noisy objects.** We verify the effectiveness of measured data uncertainty in filtering out noisy samples from the training set. Table 3 reports the results of discarding samples corresponding to the highest 5% and 10% uncertainty scores (i.e., filtering out possible noisy samples) for different models. We can observe that the predictive performance of each model is improved when samples with high uncertainty scores are abandoned both for 95% data and 90% data settings, which indicates that the reliability of detecting noisy samples in the training data and these samples do not contribute valuable supervision to model training. Therefore, our data uncertainty scores can serve as effective indicators for identifying noisy samples and mitigating the model learning from misleading supervisory information, thereby enhancing predictive performance.

**Filtering of redundant objects.** Afterwards, we examine the effectiveness of the redundant samples filtering by comparing uncertainty-aware and random discarding (i.e., uniformly dropping a certain

Table 3: Performance of filtering out noisy samples using uncertainty scores on COCO 2017 valset. "95%" represents retaining samples less than 95% quantile of aleatoric uncertainty scores.

| Model | Data(%) | AR | AP | $AP_{50}$ | $AP_{75}$ | $AP_L$ | $AP_M$ | $AP_S$ |
|---|---|---|---|---|---|---|---|---|
| YOLOX-S | 100 | 53.92 | 39.43 | 57.62 | 42.53 | 52.53 | 43.24 | 21.22 |
| | 95 | **54.34** | **39.78** | **58.36** | 42.77 | 51.97 | **43.85** | **22.56** |
| | 90 | 53.77 | 39.41 | 58.15 | **42.91** | 51.87 | 43.55 | 21.43 |
| YOLOX-M | 100 | 56.98 | 44.05 | 61.72 | 47.44 | 58.33 | 48.32 | 26.65 |
| | 95 | **58.62** | **44.86** | 63.17 | 48.54 | 58.39 | 49.15 | **27.26** |
| | 90 | 58.44 | 44.84 | **63.22** | **48.61** | **58.72** | **49.51** | 26.54 |
| Deformable DETR | 100 | 67.44 | 46.22 | 65.23 | 50.00 | 61.73 | 49.21 | 28.82 |
| | 95 | **69.52** | **47.31** | **67.14** | **51.24** | 62.55 | **50.53** | 29.73 |
| | 90 | 69.37 | 47.22 | 67.02 | 51.06 | **62.76** | 50.24 | **29.81** |

Table 4: Performance of filtering out redundant samples using uncertainty-aware filter and uniform sampling on COCO 2017 valset. "95%" represents abandoning 5% redundant samples.

| Model | Data(%) | AR | AP | $AP_{50}$ | $AP_{75}$ |
|---|---|---|---|---|---|
| | | | Random / Ours | | |
| YOLOX-S | 95 | 53.92/**54.05** | 37.72/**39.54** | 54.68/**58.37** | 40.66/**43.12** |
| | 90 | 53.85/**53.86** | 36.74/**39.12** | 52.94/**57.88** | 39.83/**42.44** |
| | 80 | 53.51/**53.56** | 36.46/**39.05** | 52.81/**57.94** | 39.59/**42.13** |
| | 70 | 53.15/**53.21** | 35.12/**38.66** | 53.83/**57.53** | 38.18/**41.44** |
| YOLOX-M | 95 | 57.11/**58.68** | 43.22/**45.21** | 60.03/**63.62** | 47.05/**48.95** |
| | 90 | 57.65/**58.03** | 43.04/**44.32** | 60.12/**62.97** | 46.88/**48.21** |
| | 80 | 57.24/**57.44** | 43.01/**44.01** | 60.03/**62.23** | 46.80/**47.32** |
| | 70 | 59.25/**59.54** | 42.52/**43.66** | 60.01/**61.11** | 46.44/**46.65** |
| Deformable DETR | 95 | 68.03/**68.36** | 46.01/**46.40** | 65.12/**65.76** | 49.85/**50.33** |
| | 90 | 67.51/**68.27** | 45.59/**46.05** | 64.06/**65.99** | 49.13/**49.94** |
| | 80 | 66.24/**67.49** | 44.26/**45.47** | 63.08/**65.44** | 48.13/**49.07** |
| | 70 | 65.47/**66.71** | 43.75/**44.82** | 62.30/**64.25** | 47.27/**48.20** |

percentage of samples) strategies, with the experimental results summarized in Table 4. As shown, the proposed uncertainty-aware filtering strategy consistently outperforms uniform sampling for all metrics under different data percentages, suggesting that leveraging data uncertainty scores to cluster samples (i.e., grouping overall training data into multiple subsets with similar patterns) is reliable. Furthermore, uniforming data selection can dramatically degrade predictive performance on relatively small-capacity models like YOLOX-S. Oppositely, our uncertainty-aware data sampling still maintains superior performance, with only a marginal reduction of 0.8% in AP while discarding 30% of the data. Interestingly, uncertainty-aware data sampling with 95% data surpasses predictive performance with 100% data, which further verifies the existence of noisy samples in training data. In the future, the proposed uncertainty-aware filtering could serve as a new paradigm for data pruning.

### 5.3 Ablation studies

This section further examines the effectiveness of our aleatoric uncertainty measure on the self-driving dataset. We also conduct ablation studies on different vision backbones (e.g., DINOv2), the hyper-parameters $\beta$ in Eqn. 9 and the combination of aleatoric uncertainty-aware filter and regularizer.

Table 5: Results (AP) on the self-driving dataset: BDD100K. "95% data" denotes abandoning 5% samples with the highest uncertainty scores, and the same meaning goes for "90% data".

| | Vanilla | UA-entropy | 95% data | 90% data |
|---|---|---|---|---|
| YOLOX-S | 28.16 | **32.53** | **33.41** | **33.38** |
| YOLOX-M | 30.17 | **34.02** | **34.15** | **34.20** |
| D-DETR | 65.33 | **68.72** | **69.02** | **68.81** |

**Effectiveness on the self-driving dataset.** We further examine the performance of the proposed data uncertainty measure for object detection in the self-driving scenario using the BDD100K dataset Yu et al. (2020). This large-scale and long-tailed driving video dataset includes a diverse

set of 100k annotated images (70k/10k/20k images for train/val/test set) with 10 classes for object detection. Table 5 presents the performance of data uncertainty scores used for entropy regularization and noisy sample filtering, showing significant gains in terms of average precision (AP) on YOLOX and Deformable DETR. We can especially observe a more prominent gain on small-scale YOLOX-S. All of this further confirms the superior scalability of our method across different real-world tasks.

**Different vision backbones.** We further examine the applicability of our approach to additional vision foundation models, such as SAM2 and DINOv2 Oquab et al. (2023). Due to the inherent resolution limitation of DINOv2, we incorporate LoftUp Huang et al. (2025) (Learnable Feature Upsampling)—a recent technique designed to enhance the spatial resolution of features extracted from vision backbones—to upsample its feature maps before applying our framework. Table 6 reports the performance of UA-entropy with DINOv2 enhanced by LoftUp as well as SAM2 on the COCO 2017 val set. As shown, our method consistently improves performance across both backbones, further demonstrating its generalizability and effectiveness.

Table 6: Performance of UA-entropy with DINOv2 and SAM2 on COCO 2017 valset.

| Model | Method | DINOv2 | | SAM2 | |
|---|---|---|---|---|---|
| | | AR | AP | AR | AP |
| YOLOX-S | Vanilla | 53.91 | 39.43 | 53.91 | 39.43 |
| | UA-Entropy | **54.31** | **39.83** | **54.49** | **39.85** |
| YOLOX-M | Vanilla | 57.92 | 44.34 | 57.92 | 44.34 |
| | UA-Entropy | **58.90** | **45.39** | **58.96** | **45.45** |

**Regularization coefficient $\beta$.** We analyze the effect of hyper-parameter $\beta$ on predictive performance in the loss function 9. Table A2 reports the comparison results under various $\beta$ for constant weighting and uncertainty-aware entropy regularization on YOLOX-S. As shown, the entropy penalty with a constant weighting is particularly sensitive to hyper-parameter $\beta$, with large values (e.g., 0.4) resulting in significantly poor performance. In contrast, the proposed data uncertainty-aware regularizer is robust to $\beta$ owing to sample-uncertainty adaptive weighting, which highlights that data uncertainty provides a more reliable way to balance difficult and easy samples.

**Combination of uncertainty-aware data filter and regularization.** Table 2 and 4 have shown the effectiveness of data uncertainty for redundant sample filtering and entropy regularization. It is worth exploring whether the performance of uncertainty-aware data filtering can be further enhanced by incorporating sample-adaptive regularization. To this end, we compare the predictive performance of using uncertainty-aware data filtering alone versus its combination with sample-adaptive regularization under different proportions of training data. As shown in Fig. A4 in Appendix, the proposed sample-adaptive regularization (UA-entropy) consistently improves the performance of redundant sample filtering on YOLOX-S and YOLOX-M by incorporating data uncertainty of each object to adaptively balance the impact of easy and hard samples within the remaining data.

## 6 Conclusions

This work investigates an important yet under-explored problem – how to accurately characterize aleatoric uncertainty in object detection. Profiting from the powerful feature representation capabilities of vision foundation models, we propose to estimate the aleatoric uncertainty of each object based on the representation space of foundation models. Furthermore, we explore two practical uncertainty-related tasks: aleatoric uncertainty-aware sample filtering and loss regularization. These tasks serve a dual purpose: examining the quality of aleatoric uncertainty and being used to develop a data-centric learning paradigm aimed at enhancing model performance and training efficiency. Extensive empirical studies validate the effectiveness of the proposed aleatoric uncertainty measure, demonstrating consistent performance gains across various advanced detection models.

In the future, we can explore leveraging various vision foundation models, e.g., DINOv2 Oquab et al. (2023) and GroundingDINO Liu et al. (2023b), to quantify data uncertainty at the detection level. Additionally, it is critical to develop more techniques, such as knowledge distillation, to extract valuable knowledge from foundation models for uncertainty quantification. Large Vision Language Models (LVLMs) Liu et al. (2024); Zhu et al. (2023) bridge the gap between visual and linguistic understanding and exhibit the potential towards achieving general artificial intelligence. However, they also easily produce hallucinations or generate inconsistent responses with input images Liu et al. (2023a); Zhou et al. (2023); Li et al. (2023b). Typically, LVLMs are fine-tuned on language-image instruction-following data generated from COCO, so the proposed noisy sample filtering strategy could be beneficial in enhancing robustness and mitigating hallucinations.

**Reproducibility Statement** We have made significant efforts to ensure the reproducibility of our work. The details of model architectures, training settings, and evaluation protocols are provided in the main paper. Additional implementation details, hyperparameter configurations, and ablation results are included in the appendix. Furthermore, the complete source code and scripts necessary to reproduce our experiments are provided in the supplementary materials.

## REFERENCES


Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*, 2023.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 213–229. Springer, 2020.

Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5710–5719, 2020.

Peng Cui, Yang Yue, Zhijie Deng, and Jun Zhu. Confidence-based reliable learning under dual noises. In *Advances in Neural Information Processing Systems*, 2022.

Peng Cui, Dan Zhang, Zhijie Deng, Yinpeng Dong, and Jun Zhu. Learning sample difficulty from pre-trained models for reliable prediction. *Advances in Neural Information Processing Systems*, 36, 2024.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pp. 1184–1193. PMLR, 2018.

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

Thomas G. Dietterich and Alex Guyer. The familiarity hypothesis: Explaining the behavior of deep open set methods. *Pattern Recognition*, 132:108931, December 2022. ISSN 0031-3203. doi: 10.1016/j.patcog.2022.108931. URL http://dx.doi.org/10.1016/j.patcog.2022.108931.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, January 2015.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059, 2016.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Haiwen Huang, Anpei Chen, Volodymyr Havrylov, Andreas Geiger, and Dan Zhang. Loftup: Learning a coordinate-based feature upsampler for vision foundation models, 2025. URL https://arxiv.org/abs/2504.14032.

Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. Open-set image tagging with multi-grained text supervision. *arXiv e-prints*, pp. arXiv–2310, 2023a.

Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*, 2023b.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.*, 110(3):457–506, 2021. doi: 10.1007/S10994-021-05946-3.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20578–20589. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/ecb9fe2fbb99c31f567e9823e884dbec-Paper.pdf.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/li23q.html.

Shengze Li, Jianjian Cao, Peng Ye, Yuhan Ding, Chongjun Tu, and Tao Chen. Clipsam: Clip and sam collaboration for zero-shot anomaly segmentation, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023b.

Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. Gmmseg: Gaussian mixture based generative semantic segmentation models. In *Advances in Neural Information Processing Systems*, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

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

Xinyu Liu, Wuyang Li, Qiushi Yang, Baopu Li, and Yixuan Yuan. Towards robust adaptive object detection under noisy annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14207–14216, 2022.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, October 2021.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33:15288–15299, 2020.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24384–24394, 2023.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Jie Jessie Ren, Stanislav Fort, Jeremiah Zhe Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *ArXiv*, abs/2106.09022, 2021. URL `https://api.semanticscholar.org/CorpusID:235458597`.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.

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

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7262–7272, 2021.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: fully convolutional one-stage object detection. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pp. 9626–9635, 2019.

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 24261–24272, 2021.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pp. 9690–9700. PMLR, 2020.

Huang Xiaoke, Wang Jianfeng, Tang Yansong, Zhang Zheng, Hu Han, Lu Jiwen, Wang Lijuan, and Liu Zicheng. Segment and Caption Anything. In *CVPR*, 2024.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 2633–2642. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00271.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2022.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Wang Zhang, Ziwen Martin Ma, Subhro Das, Tsui-Wei Lily Weng, Alexandre Megretski, Luca Daniel, and Lam M. Nguyen. One step closer to unbiased aleatoric uncertainty estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15):16857–16864, Mar. 2024a. URL https://ojs.aaai.org/index.php/AAAI/article/view/29627.

Yichi Zhang, Zhenrong Shen, and Rushi Jiao. Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine*, 171:108238, 2024b.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2023.

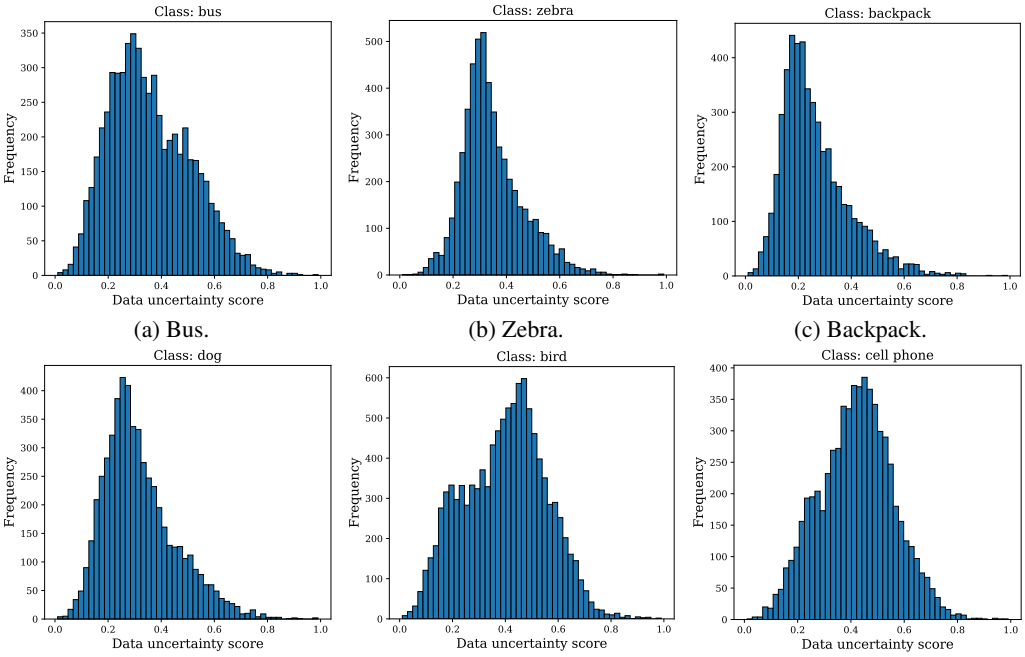
Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=gZ9hCDWe6ke.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021b.

Table A1: AP of filtering out 10% redundant samples with different bins on COCO 2017 valset.

| Models / Bins | 6 | 8 | 10 | 12 | 16 |
|---|---|---|---|---|---|
| YOLOX-S | 39.04 | 39.10 | **39.12** | **39.14** | 39.07 |
| YOLOX-M | 44.19 | 44.29 | **44.32** | 44.31 | 44.21 |
| Deformable DETR | 45.91 | 46.03 | **46.05** | 46.02 | 45.96 |

## A HISTOGRAMS OF ALEATORIC UNCERTAINTY SCORES BY CATEGORIES




(a) Bus. (b) Zebra. (c) Backpack.

(d) Dog. (e) Bird. (f) Cell phone.

Figure A1: Distributions of data uncertainty scores for different classes.

## B MORE SORTING EXAMPLES VIA ALEATORIC UNCERTAINTY ON MS-COCO

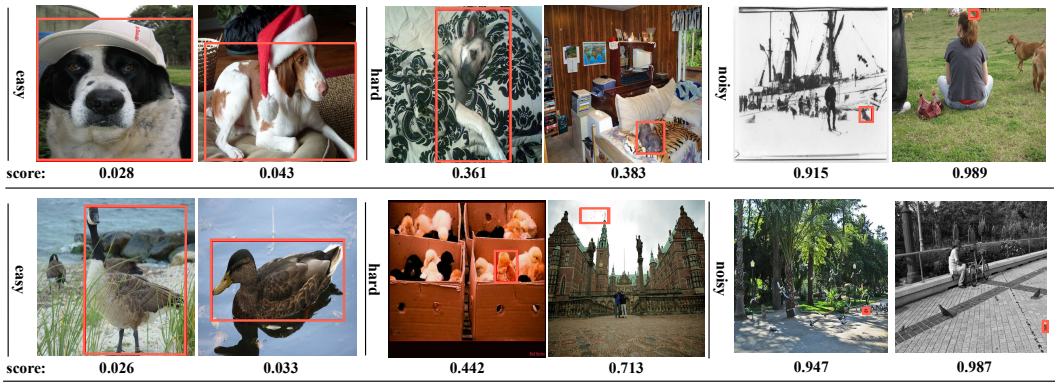



Figure A2: Visualization of scoring objects with corresponding uncertainty scores in training images of MS-COCO Lin et al. (2014) for class “dog” and “bird”. The orange bounding box is the annotated ground truth. “Hard” objects suffer from occlusion or obscureness within an image, and “noisy” ones have misleading bounding boxes.

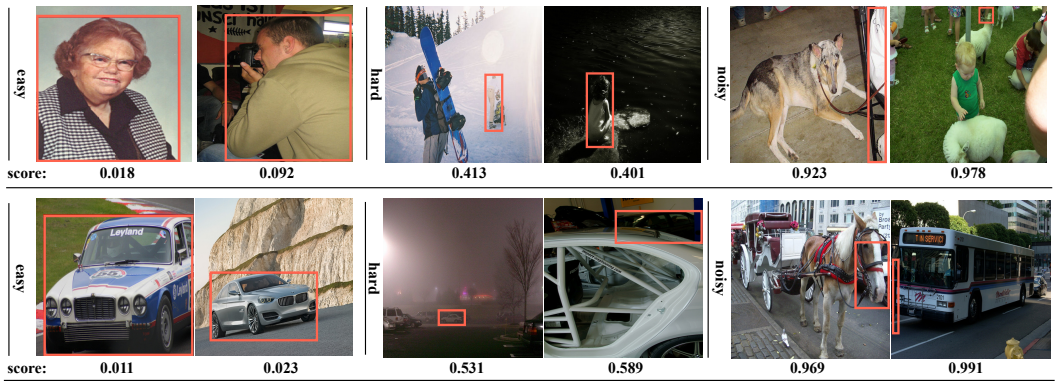


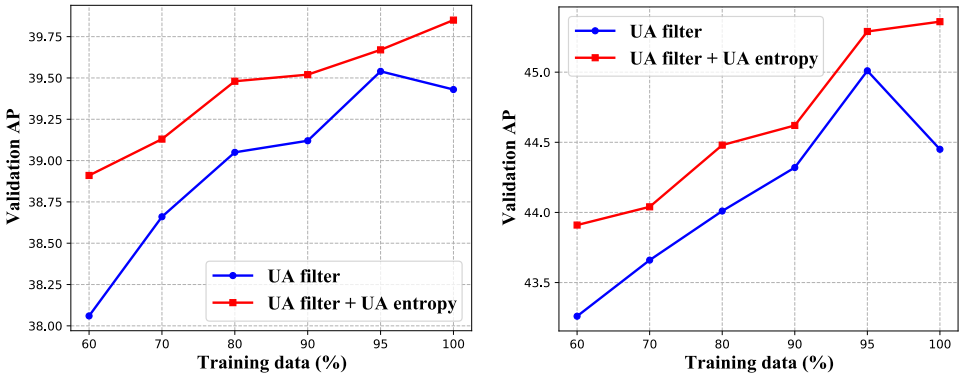

Figure A3: Visualization of scoring objects with corresponding uncertainty scores from training images of Pascal VOC Everingham et al. (2015) for class "person" and "car".

Table A2: The comparison for AP and AR under different $\beta$ on YOLOX-S. **Bold** indicates the results from the chosen hyperparameter.

| $\beta$ | | 0 | 0.10 | 0.20 | 0.25 | 0.30 | 0.40 | 0.50 |
|---|---|---|---|---|---|---|---|---|
| Entropy | AP | 39.43 | 37.22 | **38.55** | 37.75 | 37.47 | 36.85 | 36.03 |
| | AR | 53.92 | 51.88 | **52.97** | 52.10 | 51.90 | 51.02 | 50.42 |
| UA-entropy | AP | 39.43 | 39.75 | **39.85** | 39.81 | 39.77 | 39.54 | 39.44 |
| | AR | 53.92 | 54.17 | **54.26** | 54.24 | 54.02 | 53.88 | 53.91 |

Table A3: Performance comparison of uncertainty-aware regularizer (UA-entropy) and constant entropy regularizer (Entropy) on COCO 2017 valset.

| Model | Method | AR | AP | $AP_{50}$ | $AP_{75}$ | $AP_L$ | $AP_M$ | $AP_S$ |
|---|---|---|---|---|---|---|---|---|
| FCOS | Vanilla | 57.21 | 41.46 | 60.71 | 45.08 | 51.53 | 44.82 | 24.41 |
| | Entropy | 57.10 | 41.35 | 60.62 | 45.01 | 51.61 | 44.24 | 24.03 |
| | UA-entropy | **58.62** | **42.62** | **62.13** | **46.36** | **52.74** | **45.67** | **25.33** |
| DINO | Vanilla | 72.63 | 49.39 | 66.97 | 53.84 | 63.64 | 52.30 | 32.48 |
| | Entropy | 72.23 | 49.30 | 66.78 | 53.11 | 63.06 | 51.65 | 32.15 |
| | UA-entropy | **73.59** | **49.59** | **67.01** | **54.23** | **63.67** | **53.07** | **32.54** |




(a) The performance comparison on YOLOX-S. (b) The performance comparison on YOLOX-M.

Figure A4: The effect of combining uncertainty-aware redundant samples filtering (UA-filter) and regularization (UA-entropy) on performance on YOLOX-S and YOLOX-M.