Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 3. 2025**

Article's History: Received: 28.08.2025 Revised: 14.11.2025 Accepted: 23.12.2025

UDC 004.85 DOI: 10.31649/vitce/3.2025.54

Active self-learning for object detection in an imbalanced data environment: The TAAST approach

Dmytro Ivanov*

Posrgraduate Student
Zhytomyr Polytechnic State University
10005, 103 Chudnivska Str., Zhytomyr, Ukraine
https://orcid.org/0000-0002-7386-4497

Abstract. In the context of the growing development and application of computer vision, there is a growing need to reduce the cost of manual data markup, especially in tasks of detecting rare objects in conditions of long-tailed class distribution. The purpose of the study was to improve the efficiency of identifying rare image categories by improving the active self-learning strategy. The study used the Tail-Aware Active Self-Training approach, which was based on strategic selection of frames, considering the entropy of uncertainty, class rarity, and semantic diversity in the feature space of the Contrastive Language-Image Pretraining model, followed by the use of pseudo-markup using the You Only Look Once detector, version 8. As a result of experiments on Large Vocabulary Instance Segmentation datasets, version 1.0, and nuImages-imbalanced, the proposed strategy provided an increase in AP_rare accuracy by 6.3-6.4 percentage points compared to the basic Random and Uncertainty Sampling approaches. The overall accuracy of the model did not decrease, but increased to 36.0-43.2% mAP, depending on the dataset. The markup efficiency indicator reached 42-43%, which was 9-10 points higher than competitive strategies. The results of the experiment were statistically reliable, since the confidence intervals for the AP_rare accuracy metric in the case of using the Tail-Aware Active Self-Training method do not overlap with the intervals for the basic random and Uncertainty-only strategies. This indicated that the advantage of this method was not random, but was confirmed with high probability. Consequently, the results obtained demonstrated the reliability and stability of the proposed approach. It was demonstrated that after two active iterations, the model reached a performance plateau, which significantly reduced computational costs. The practical significance of the study lies in creating an effective tool for automated deployment of computer vision models in conditions of a limited markup budget

Keywords: machine learning; semantic clustering; pseudoanotation; entropy sampling; class balancing; computer vision; markup optimisation

Introduction

Computer vision systems are rapidly developing, and advanced object detection models demonstrate high accuracy on balanced data sets. However, in real-world problems, images often have a long-tailed distribution: most objects belong to categories with low representation in the sample. Under such conditions, models lose their ability to effectively generalise to rare classes, which is critical for applications in biomonitoring, autonomous driving, safety, etc. This problem is becoming particularly relevant due to the growing need for automated data processing in high-risk or hard-to-reach environments, where markup for a large number of images is extremely resource-intensive.

Recent studies confirm that the main reason for the low efficiency of object detection models in rare classes is a pronounced class imbalance in image sets. In particular, in LVIS (Large Vocabulary Instance Segmentation), nuImages, and iNaturalist, most classes have less than 10 examples, which reduces the quality of recognition and is masked by the global mean Average Precision (mAP) metric. As noted by Y. Li *et al.* (2020), the distribution of objects in the LVIS v1.0 Set obeys Zipf's law: only a third of classes have more than 100 examples, and more than 28% have less than 10. A similar situation was described by H. Caesar *et al.* (2020) for the nuImages dataset designed for realistic autonomous driving scenes – more than half of the classes occur less than 10

Suggested Citation:
Ivanov, D. (2025). Active self-learning for object detection in an imbalanced data environment: The TAAST approach. *Information Technologies and Computer Engineering*, 22(3), 54-64. doi: 10.31649/vitce/3.2025.54

*Corresponding author

times. In an analysis of the iNaturalist dataset for biodiversity tasks (De Alvis & Seneviratne, 2024), the ratio between common and rare classes exceeds 1:500, which reduces the accuracy of the latter by almost 20 percentage points. Such results show that a global metric such as mAP can mask critically poor quality for underrepresented categories.

To mitigate the impact of the imbalance in long-tailed object detection problems, the researchers proposed a number of modifications to the loss functions and classification layers. The Balanced Group Softmax method (Li *et al.*, 2020) implements group normalisation of logits in accordance with the frequency of classes, which provides an increase in ap_rare accuracy by 4-5 percentage points. Within the unified Balanced Classification scheme (Qi *et al.*, 2023) an approach to loss weighting is generalised, allowing adaptation to different degrees of class representation. The method developed by B. Li *et al.* (2022) with multimodal learning using pseudo labels at the image level also demonstrates improved accuracy on low-frequency classes. C.-L. Duan *et al.* (2024) presented the DRCL-2 approach, which combines contrast training with the reconstruction task and helps to further increase AP_rare by 5 percentage points.

However, the above methods are model-oriented, i.e., they involve complete markup of data sets, which does not solve the problem of high time and resource costs. In this context, active learning is promising, where the model requests annotations only for the most informative samples. Y. Gal & Z. Ghahramani (2016) proposed Entropy Sampling, a strategy for selecting images with maximum forecast entropy. The CoreSet Sampling approach, presented by O. Sener & S. Savarese (2018), provides sample diversity through clustering in feature space. J. Wu *et al.* (2022) integrated these ideas into the Entropy-AL + Progressive Diversity method, which increased mAP by 3 percentage points within a fixed budget on the COCO (Common Objects in Context) set. As indicated by K. Sohn *et al.* (2020) and M. Xu *et al.* (2021), most active strategies do not consider class rarity – the annotation budget is often spent on already well-represented categories, while rare classes are ignored.

Thus, existing approaches are either aimed at improving accuracy in rare classes without reducing markup costs, or optimise the annotation budget without considering class rarity. The lack of methods that combine both approaches – focusing on rare classes and saving manual markup – creates a noticeable gap in current research. The purpose of this study was to develop and empirically test an active self-learning strategy for the object detection problem focused on rare classes under long-tailed distribution conditions. To achieve this goal, the following is proposed: sampling by uncertainty, weighted by class frequency; clustering to ensure diversity in the space of semantic features obtained using CLIP (Contrastive Language-Image Pre-training); pseudo-labelling with a confidence threshold of 0.8. The study's hypothesis was that this approach would reduce the amount of manual annotation while maintaining or even improving accuracy in rare classes.

Materials and Methods

The developed Tail-Aware Active Self-Training (TAAST) approach is presented as a formalised technique that covers the full cycle of active self-learning training of the object detection model – from using the initial seed set and pseudo-marking to selecting informative examples and further training of the detector. Formalisation in the form of an algorithm and optimisation problem guarantees reproducibility of experiments and provides a reasonable assessment of the effectiveness of the strategy. Formalisation of the cycle of the active-self-learning process of training a model for the problem of object detection in conditions of limited access to annotated data was carried out according to a typical practical scenario, in which:

- a large array of unanotated images U is available (for example, from cameras of driver assistance systems – Advanced Driver-Assistance Systems (ADAS); aerial photos from drones; log files of multi-season monitoring, etc.);
- a small initial set of annotated data L is available, covering approximately 10% of the total volume – such a "seed set" is usually formed as part of the pilot stage;
- fixed budget B of frames allocated for each iteration of active learning;
- total number of active iterations is denoted as T.

To maximise the accuracy of the model in rare classes with the minimum possible volume of new annotated examples, it was proposed to integrate active learning with a self-learning approach, where the model used its own predictions to expand the learning set. The target metric for evaluating the effectiveness of the proposed method was the average accuracy value calculated separately for rare classes (mAP_rare). This helped to focus attention on exactly the subset of objects that is traditionally most vulnerable to imbalances and lack of training examples.

The YOLOv8-s model (version 8, small configuration) was chosen as the basis for the system, which demonstrates a sufficient level of accuracy (~38% mAP) on the COCO set with a significantly lower computational load compared to larger variants (Jocher *et al.*, 2023). A special feature of this architecture is the use of an anchor-free detection head, which does not require preliminary determination of object sizes and better summarises examples that rarely occur in the training set, in particular, on the "long tail" of the distribution (Tian *et al.*, 2019). Based on initialisation with weights previously trained in COCO, the model is already able to generate fairly accurate pseudo-labels in the first active cycle without additional configuration. To calculate the semantic similarity of scenes, the CLIP model with the VIT-L/14 (Vision Transformer) architecture was used, which was trained on paired text – image examples and can encode plot features in the form of compact 512-dimensional vectors (Radford *et al.*, 2021). Clustering of these vectors was performed using an algorithm k-means++, which provided fast and stable splitting of a large number of vectors into groups due to improved centre initialisation (Johnson *et al.*, 2021). After filtering by value and diversity, the frames were grouped into B clusters (by the number of

available signatures), and the representative frame closest to the centre was selected from each cluster. The model was further trained on a combined set that contained initial human annotations, new marked frames, and pseudo-annotations with high confidence (⩾ 80). Optimisation was performed using stochastic gradient descent (SGD) with a momentum of 0.937 and a regularisation coefficient (weight decay) of $5 \cdot 10^{-45}$, one of the most effective optimisation methods in machine learning problems (Bottou, 2012). The warm-up (initial phase) lasted 300 epochs with a linear decrease in the learning rate from 0.01 to 0.001, while each subsequent active cycle covered only 30 epochs, which helped to quickly adapt the model to new data without overtraining.

All experimental studies were conducted on two commonly used datasets: LVIS v1.0 and nuImages-imbalanced (an imbalanced version of the nuScenes subset). For each of them, a controlled division scheme was applied into training and validation parts. In particular, 10% of the data was randomly selected from the initial training sample to form the initial body with manual marking, which is further designated as L_0 (seed-dataset). The remaining 90% of the training images formed a U pool that simulated a realistic situation of incomplete markup when starting a new data collection project. This distribution allows simulating the conditions of a limited human resource at the beginning of active training.

In this paper, a class was considered rare if it was found in less than 10 examples in the initial training set, which meets the LVIS-taxonomy criteria (Li *et al.*, 2022). To objectively evaluate the performance of the model, validation subsets were used, which remained fixed throughout all stages of the experiment. Evaluation of the test sample was performed only once – after all active cycles were completed, to avoid information leakage and re-evaluation of the results.

All active learning strategies in the study were implemented in three consecutive iterations ($T = 3$), which was chosen empirically: in two cycles, the potential of rare classes was not yet exhausted, while after the fourth cycle, the increase in the average accuracy metric for rare classes became less than 0.3 percentage points. In each cycle, the model generated a pool of pseudo-annotated examples P, adding to it all the provided objects for which the model confidence level exceeded the threshold of 0.8. Next, the top-20% filter was used for the integral significance score (x), which allowed excluding examples with too low a value and reduce duplication of head scenes. Semantic clustering was performed in this upper quintile subset, and one representative frame was selected from each cluster, for a total of $B = 256$ images per cycle. After each active cycle, the model was further trained on the combined set with $L \cup Q \cup P$ for 30 epochs using stochastic gradient descent and cosine reduction of the learning rate (from 0.01 to 0.001), in accordance with the recommendations for YOLOv8 (Jocher *et al.*, 2023). The same set of hyperparameters and procedure was used for all experimental strategies, which ensured the purity of comparison and made implicit reconfiguration impossible. As part of the experimental study, a consistent comparison of Random → Uncertainty-only → TAAST strategies was performed.

Results and Discussion

Stages of implementing the Tail-Aware Active Self-Training strategy

A detailed description of the sequence of actions within one active cycle of active self-learning of the object detection model ensures the reproducibility of the experiment. In addition, it helps to clearly understand the proposed method and evaluate its effectiveness. Below is a step-by-step scheme for implementing the TAAST strategy, where each step reflects the logic of the transition from automatic pseudo-markup generation to an optimisation training goal.

Step 1. Pseudo-markup based on confidence forecasts

The first stage of the proposed active self-learning strategy was to automatically expand the training set using the most reliable model predictions. This approach helped to reduce the amount of manual marking, while maintaining the quality of the training signal. For each object detected by the model in the unsigned image pool U, the level of trust in the object's belonging to a certain class was calculated. The object was moved to a set of pseudo-labels P, if its highest predicted probability exceeded a pre-determined confidence threshold. This was formalised by the following equation (proper formulation):

$$p_{max} = \max_k p_k \geq \tau, \tau = 0{,}8, \quad (1)$$

where p_k – probability that the detected object belongs to k-th class; $\max_k$ – maximum probability among all classes, i.e., the model's confidence in the most probable hypothesis; τ – confidence threshold is set at 0.8 (or 80%).

Selection of a threshold value $\tau = 0{,}8$ was based on previous experiments, where it was found that this level of trust provides an optimal compromise between the number of examples added and the noise level in pseudo-marking. Too low values of τ lead to a large number of false examples, while too high ones reduce the effectiveness of increasing the training set due to a limited number of confident forecasts.

Step 2. Assessment of frame value by uncertainty

After the most reliable predictions of the model are transferred to the pseudo – markup set, the next step is to evaluate the value of the remaining images from the unlabelled pool. It is necessary to select those examples that, when labelled manually, will bring the greatest increase in accuracy. The main criteria for such an assessment are the uncertainty of the model in relation to a particular frame, the presence of rare classes, and its diversity in the context of the entire sample. To quantify the uncertainty of the model with respect to the image , the sum of entropies is used for all objects detected in this image. In particular, the calculation is performed using the equation (adapted from C.-L. Duan *et al.* (2024)):

$$E(x) = \sum_{b \in \hat{y}(x)} \left(-\sum_{k=1}^{C} p_{k,b} \log p_{k,b} \right), \quad (2)$$

where $\hat{y}(x)$ – set of all provided objects (boxes) in the image x obtained from the model; b – separate box, i.e., a rectangular area that corresponds to the detected object; $p_{k,b}$ – probability that box b belongs to k-th class; C – total number of classes.

Entropy, as a measure of uncertainty, increases when the probability distribution is "flat", meaning that the model does not have a clear advantage in favour of any class. Accordingly, a high value of $E(x)$ indicates the complexity of the image for the model and the feasibility of marking it manually. This approach allows focusing limited resources on those images that can significantly improve the training of the model in the active cycle.

Definition of a candidate Image class in active learning

After estimating the overall uncertainty, it is necessary to determine the class of objects for which the model shows the greatest confusion in a particular image. To do this, all the predicted objects (frames, or boxes) on the frame are analysed, and the one in which the model has the lowest overall confidence is selected – that is, even the highest probability of belonging to any class is low. This allows identifying the "weakest point" for the model in a given image and the corresponding class as a candidate for improvement using manual annotation. Formally, this process is defined as follows (proper wording):

$$b^* = \arg \min_{b \in \hat{y}(x)} \left(\max_k p_{k,b} \right), \quad (3)$$

where x – image being analysed; $\hat{y}(x)$ – set of all provided objects (boxes) in the image x; $b \in \hat{y}(x)$ – specific frame within this image; $p_{k,b}$ – probability that the frame b belongs to the class k; $\max_k p_{k,b}$ – highest probability, which reflects the model's confidence level in its forecast for the frame b.

Thus, b^* indicates the frame for which the model is least confident, even in terms of its strongest prediction. Further, for this frame, the so-called candidate image class is defined – the class that the model still considers most likely for the frame b^* (actual wording):

$$c(x) = \arg \max_k p_{k,b^*}. \quad (4)$$

Class $c(x)$ is considered a representative of the category that the model confuses most in the image x. If the detected candidate class belongs to rare categories, the image gets a higher priority for subsequent manual markup as part of active learning. This allows effectively using a limited annotation resource, focusing it on examples that help to improve the accuracy of the model on rare classes.

Evaluation of the current representation of a class to determine its rarity

The next step is to evaluate how well the candidate class is represented $c(x)$, which caused the most uncertainty in the model in the current training set. To do this, the number of available examples of this class is calculated considering both manually annotated and pseudo-labelled samples. Evaluation is performed using the following expression (author's wording):

$$n_{c(x)} = |L_c(x)| + |P_{c(x)}|, \quad (5)$$

where $c(x)$ – class of object that the model considers most likely in the most unreliable area of the image x; $L_c(x)$ – subset of manually marked-up class images $c(x)$, included in the training set L; $P_{c(x)}$ – subset of class images $c(x)$ that were automatically added as pseudo-markings to the set P; $|\cdot|$ – operator that defines the number of elements in a set.

This equation allows quantifying the "saturation" of an individual class in the current data set. Low value of $n_{c(x)}$ indicates that the corresponding class is still rare, and new examples involving it may be of high value in the context of active selection. A high value means that the class is already sufficiently represented, and additional marking of frames with its presence is less of a priority.

Calculation of the frequency weight for rare classes in the sample

In order to give preference to images that contain rare categories when ranking examples, the number of available examples of the class $n_{c(x)}$ is converted to a weighting factor that is inversely dependent on the frequency of this class. This weight is determined by the following equation (author's wording):

$$\omega_{c(x)} = \frac{1}{\log(n_{c(x)} + \beta)} \quad (6)$$

where $\omega_{c(x)}$ – weighting factor for the class clas $c(x)s$, which is used later to prioritise the frame; $n_{c(x)}$ – total number of class images $c(x)$ available in the training set (both manually annotated and obtained as a result of pseudo-markup), calculated according to equation (5); β – positive constant that guarantees the certainty of a logarithmic function even when the number of examples of the class is zero; in this paper, the value is assumed to be $\beta = 1$.

This equation allows compensating for bias in favour of frequently presented classes. Due to logarithmic smoothing of values, the weighting factor $\omega_{c(x)}$ increases for classes with few examples and decreases for well-represented classes. Thus, even on a limited budget, active training with high priority selects those frames that can improve the accuracy of the model in poorly represented categories.

Calculation of the integral value of a frame for further example selection

To make an effective comparison between all images left without annotations, it is proposed to combine two previously calculated characteristics – the uncertainty of the model with respect to the image and the frequency weight of the associated class – into a single integral indicator. This indicator is determined by the equation (author's wording):

$$\varphi(x) = \omega_{c(x)} \cdot E(x), \quad (7)$$

where $\varphi(x)$ – total (integral) value of the image , which reflects its importance for further markup; $\omega_{c(x)}$ – class

frequency weight $c(x)$, to which the model gives the highest (but not yet certain enough) probability; this coefficient is calculated by equation (6) and is higher for rare classes; $E(x)$ – total entropy of all detected objects in the image x, which characterises the degree of uncertainty of the model.

This equation allows combining the semantic complexity of the frame (due to entropy) with information about the relevance of the class (due to frequency weight), which makes it an effective criterion for selecting examples for manual annotation. Therefore, all unmarked images are sorted by value $\varphi(x)$, and the frames with the highest values are selected for manual markup. This approach allows allocating a limited annotation budget to the most valuable examples for training the model.

Step 3. Selection of different frames based on semantic diversity

After each unsigned image , an integral value was assigned $\varphi(x)$ (equation 7), it is necessary to select the frames that will most contribute to improving the accuracy of the model. These are images that have a high potential for information content and help to cover rare categories of objects. Since manual markup has a limited budget, it is marked as B (number of images that can be annotated at each iteration), it is important not only to identify the most valuable samples from the standpoint of integral metrics, but also to ensure their diversity.

The selection process is implemented in two stages:

1. Filtering by value – all images are sorted in descending order by function value $\varphi(x)$, after which a preliminary pool of candidates is formed, which includes $B' > B$ examples with the highest scores. Value B' is set empirically (for example, within 2-3 × of B), to provide sufficient space for the next step – diversification.

2. Ensuring diversity – to avoid excessive repetition of similar scenes or classes among the selected samples, a clustering mechanism is applied in the feature space. This study utilised embeddings obtained using a pre-trained CLIP model. Images from the previous pool are grouped using an algorithm of k-means, and the closest representative to the centroid is selected from each cluster. This forms the final set of B-images that will be submitted for manual marking.

This approach allows combining information content (high values $\varphi(x)$ with a variety of samples. This is crucial to ensure generalisability of the model and avoid over-training it on too uniform examples. In addition, it makes optimal use of the limited resources of human annotation within the active self-learning cycle.

Pre-filtering by integral value

At the first stage of selecting images for manual marking, a preliminary cut-off of unpromising frames was performed. This allows focusing computing resources and human attention on the most informative examples and thereby increasing the effectiveness of active learning. To form the previous set of priority examples, a subset is defined S_φ:

$$S_\varphi = \{x \in U \setminus P \mid \varphi(x) \in \text{топ} - 20\%\}, \quad (8)$$

where U – multiple of all unassigned images; P – subset of images that are already included in the pseudo-markup set; $E(x)$ – total entropy of all detected objects (frames) in the image x, which characterises the degree of uncertainty of the model; $x \in U \setminus P$ – images that remain unsigned and were not automatically annotated; $\varphi(x)$ – integral value of the frame , calculated by equation 7; S_φ – subset of the highest priority images included in the top 20% by value $\varphi(x)$.

This procedure generates many examples that are potentially most useful for manual annotation, since they combine high model uncertainty and belonging to rare classes. This approach allows reducing computational costs and using the annotation budget more efficiently, avoiding the cost of insignificant snapshots. Validity of choosing a threshold value $p = 20\%$ is confirmed by the results of previous research in the field of active learning for object detection tasks, in particular, in the papers Entropy + Progressive Diversity (Wu *et al.*, 2022) and SoftTeacher-AL (Xu *et al.*, 2021), where it is recommended to use filtering in the range of 15-25% of the most valuable examples.

Transition to the space of semantic features.

At this stage, each image x from the set S_φ is converted to a compact numerical representation – a feature vector that preserves the semantic content of the image. The purpose of this transformation is to provide a space structure in which similar frames are located close to each other, and dissimilar frames are located at a greater distance. This allows effectively applying grouping methods, in particular clustering:

$$z(x) = \frac{f_{CLIP}(x)}{\|f_{CLIP}(x)\|_2} \in R^{512}, \quad (9)$$

where x – images from the set S_φ, which is pre-selected as a set of valuable frames (equation 8); $f_{CLIP}(x)$ – 512-dimensional feature vector obtained using a pre-trained CLIP model (Radford *et al.*, 2021), which displays the content of the image; $\|f_{CLIP}(x)\|_2$ – L2-norm (Euclidean length) of the feature vector; $z(x) \in R^{512}$ – normalised vector in the 512-dimensional feature space, which is used as input for further clustering, for example by the k-means method.

This mapping of semantic features into the space allows each image to match a unified numerical representation, or conditionally – its "digital DNA". This makes it easier to analyse similarities between scenes and avoids duplication when selecting images for manual markup. The use of normalised vectors ensures that clustering will be based solely on directions in the feature space, and not on absolute values of components, which is especially important when using cosine distance-based metrics.

K-mean clustering: Detection of typical scenes

After all selected images have been converted to normalised feature vectors using the CLIP model, each image x gets a vector representation $z(x) \in R^{512}$, which is placed in a common semantic space. In this space, scenes that are similar in content have close coordinates. The next step is to divide the set of these vectors into B clusters – this is exactly how many examples are planned to be submitted for manual annotation in the current active learning cycle. The

classical algorithm is used for this purpose *k*-means with improved centroid initialisation using the *k*-means++ method (Radford *et al.*, 2021). Mathematically, the problem is formulated as minimising the total square of the Euclidean distance between vectors and cluster centres (adapted from A. Radford *et al*. (2021)):

$$\min_{\mu_1,\ldots,\mu_k} \sum_{j=1}^{k} \sum_{x \in C_j} \left|\left| z(x) - \mu_j \right|\right|_2^2, k = B, \quad (10)$$

where B – number of clusters (equal to the frame budget for markup); $z(x) \in R^{512}$ – normalised feature vector obtained from the clip model for the image x; C_j – multiple images assigned to j-th cluster; $\mu_j \in R^{512}$ – centre of j-th cluster, calculated as the arithmetic mean of vectors in C_j; $||z(x) - \mu_j||_2^2$ – square of the Euclidean distance between the image vector and the centre of the cluster. This approach allows creating a representative sample of frames that are very diverse in content, which reduces redundancy and increases the efficiency of manual marking.

Creating an active markup set.

After clustering semantic vectors by the *k*-means method, each cluster C_j where $j = 1,\ldots, B$ represents a group of frames that are similar in content. Next, an active set for manual markup is generated: one of the most representative examples is selected from each cluster. This approach allows ensuring maximum coverage of the content space with a fixed budget for markup. Calculating the active sub-set Q is performed according to the following equation:

$$Q = \left\{ x_j^* | x_j^* = arg \min_{x \in C_j} ||\, z(x) - \mu_j ||_2, j = 1, \ldots, B \right\}, \quad (11)$$

where C_j – j-th cluster formed as a result of the algorithm of *k*-means; all frames in the middle C_j have similar semantic features; $\mu_j \in R^{512}$ – centre of j-th cluster, calculated as the average value of vectors $z(x)$ for all $x \in C_j$; $z(x) \in R^{512}$ – normalised feature vector obtained from the CLIP model; $||\, z(x) - \mu_j ||_2$ – Euclidean distance between the frame feature vector x and the centre of the corresponding cluster; $arg \min_{x \in C_j}$ – operator that returns the frame with the smallest distance to the centre of the cluster, i.e., the most typical frame within the cluster C_j; x_j^* – frame that best represents the cluster C_j; Q – subset of B images, each of which is selected from a different cluster. Thus, the constructed set Q ensures that each markup frame represents a unique type of scene. This can significantly increase the efficiency of spending limited human resources, reducing redundancy and helping to speed up the process of self-learning the model for object detection.

Step 4. Updating and retraining the model on the combined data set

After a set of images Q marked up by experts on the previous one, manually adds annotated data, and confident pseudo-markings are collected, the stage of additional configuration of the model on the combined sample is performed. This section presents three key equations that formalise the structure of the new training set and the process of optimising the model weights. Updating of the manual dial:

$$L^{new} = L \cup Q, \quad (12)$$

where L – multiple images that were previously marked up manually; Q – multiple frames that were annotated by experts in the current iteration; L^{new} – updated manually marked-up set. Repetitions (duplicates) are automatically deleted, so each image is presented only once.

Creation of a complete training set:

$$T^{train} = L^{new} \cup P, \quad (13)$$

where P – set of pseudo-markings obtained on the basis of confidence forecasts of the model with a confidence threshold of at least 0.8; T^{train} – combined training sample that includes both human and automatically generated markings.

Model optimisation procedure (adapted from A. Radford *et al*. (2021)):

$$\theta^{(t+1)} = \theta^t - \eta \nabla_\theta L(T^{train}; \theta^t), t = 0,\ldots, e-1, \quad (14)$$

where θ^t – model parameters at the beginning of epoch t; η – learning rate, which gradually decreases from 0.01 to 0.001 according to the cosine attenuation graph; $L(T^{train}; \theta^t)$ – loss function that combines classification and regression components specific to the YOLO architecture (Ali & Zhang, 2024); $e = 30$ – number of epochs of additional training.

After completing 30 epochs, the model updates its scales to reflect new patterns, while maintaining previous knowledge. If the number of active iterations T did not reach the specified maximum, the process returns to the beginning of the cycle, in particular, to the pseudo-marking stage, which ensures the integration of active learning with self-learning.

Step 5. Statement of the optimisation goal of the active cycle for rare classes

After a detailed review of the stages of pseudo-markup, adjusted selection of examples and additional training of the model, it is necessary to formalise the target function of active learning and the corresponding restrictions. This section defines what exactly needs to be optimised and what resources contain the best strategy. The optimisation goal is to maximise recognition accuracy for rare classes after completing the entire sequence of active loops. Formally, the objective function is written as follows:

$$\max_S AP_{rare}(M_\theta^{(T)}), \quad (15)$$

where $M_\theta^{(T)}$ – detector with parameters θ after completion of T iterations of active learning; AP_{rare} – mean accuracy for rare classes only according to the LVIS v1.0 taxonomy; S – example selection strategy that considers the uncertainty, rarity, and variety of scenes in this case.

This goal is consistent with approaches in active learning, in particular, with the wording by B. Settles (2009), however, with a particular focus on rare classes. The limit on the manual markup budget is set by the inequality:

$$|L| \leq M_0 + BT, \quad (16)$$

where $|L|$ – total number of examples that were marked up manually after all cycles were completed; M_0 – initial set with manual markup (so-called seed-set), usually 10% of the full markup; B – number of images that can be signed in one cycle; T – total number of active iterations.

Thus, this condition ensures that the total amount of manual markup corresponds to the established budget. Equations (15) and (16) together form an optimisation problem with constraints: it is necessary to maximise the accuracy gain on rare classes without exceeding the available human resource. The proposed example selection strategy focused on rare categories (tail-aware sampling) demonstrates an advantage over random or purely entropy methods.

Comparison of Tail-Aware Active Self-Training strategy with other basic approaches

As part of the experimental study, the proposed Tail-Aware Active Self-Training strategy was compared with two basic approaches that reflect the lower and intermediate limits of the effectiveness of active training. The first basic scenario is Random, in which a fixed number of examples $B = 256$ are randomly selected at each iteration from a pool of unsigned images U. This approach does not consider either the level of uncertainty of the model or the frequency characteristics of classes, and therefore acts as a minimal control that allows assessing whether there is any benefit from using active learning.

The second basic option is the Uncertainty-only strategy, which is based on the classical entropy sorting approach proposed by B. Settles (2009). In this case, the images are ranked by the total entropy of the model's predictions, and the examples that the model is most uncertain about are added to the sample. However, this strategy ignores information about the imbalance in the class representation, which is especially important for objects that belong to rare categories. Ultimately, TAAST combines the entropy approach with the weight gain of rare classes (via a multiplier $\omega_{c(x)}$) and a semantic diversity mechanism based on clustering of normalised clip feature vectors. This approach helped to avoid excessive duplication of such personnel and ensured the maximum increase in information per unit of human resource. A consistent comparison of Random → Uncertainty-only → TAAST strategies illustrated the contribution of both the fact of active learning itself and the additional effect of taking into account the structure of the long tail (tail-aware logic).

Performance evaluation was carried out using the basic AP_rare metric, i.e., average accuracy only for objects with less than 10 examples in the initial training sample. The calculation was performed by the official LVIS/COCO API at 10 IOU (Intersection over Union) thresholds in the range of 0.50-0.95 in 0.05 increments, which ensured compatibility with previous studies in the field of long-tail detection. To ensure that the improvement in AP_rare is not accompanied by a degradation in overall accuracy, the mAP_overall metric was additionally recorded – the average accuracy for all classes using the same protocol. Label Efficiency (LE) was also evaluated – the percentage of manual markup saved compared to the full training set (for example, LE = 42% means using only 58% of real labels to achieve a given quality).

A complete three-cycle experiment was performed for each strategy under study ($T = 3$), where three fixed initial seed values were used: 21, 42, and 63. At the zero cycle stage ($T = 0$) the basic values of AP_rare and mAP were recorded, and then after each active cycle ($T = 1, 2, 3$) was evaluated on a validation subset. This step-by-step assessment allowed tracking the dynamics of learning and identifying at what stage the performance plateau is reached. After the third iteration was completed, the test part was opened once for the final measurement – this allowed adjusting to the test data. Average values and 95% confidence intervals were calculated using the equation:

$$\bar{x} \pm 1.96\,\sigma/\sqrt{3}, \quad (17)$$

where $\bar{x}$ – mean metric value; σ – standard deviation of three runs with different seeds.

As part of the evaluation of the effectiveness of active learning strategies, a comparative analysis of manual labour costs and computational time was carried out with a fixed budget for three active cycles of 256 frames each (a total of 768 frames on top of the initial seed set, which was 10% of the train part of the LVIS v1.0 dataset ≈ 10,000 images). Table 1 shows that the proposed TAAST strategy achieved the highest label efficiency (LE = 42%), reducing the need for manual annotation. Specifically, it retained 9% of human effort compared to Random and Unknown-only, which showed only 33% LE. In addition, TAAST demonstrated an advantage over qualitative metrics (AP_rare), proving the effectiveness of including weighting factors for rare classes and combining pseudo-markings with semantic clustering.

Table 1. Cost analysis

Strategy	Manual frames per cycle	Manual frames per 3 cycles	Label-efficiency
Random	256	768	33%
Uncertainty-only	256	768	33%
TAAST	256	768	42%

Source: compiled by the author based on the results of the experiment

As can be seen from the table, the advantage of TAAST is a combination of classified entropy, semantic clustering, and pseudo-markup, which allows not only to reduce human effort, but also to achieve higher AP_rare accuracy values in rare classes. Thus, the experimental results confirmed that TAAST provides a more economical use of the annotation budget without compromising the quality of the model, which is a key factor for deploying systems in real-world conditions with limited resources. Table 2 summarises the totals for all active learning strategies tested on LVIS v1.0 and nuImages-Imbalanced datasets. Accuracy in rare classes (AP_rare), overall quality (mAP), human markup performance (LE), and gain after three active learning cycles were evaluated.

Table 2. Analysis of various active learning strategies

Dataset / Method	AP_rare, start	AP_rare, final ± 95% CI	Δ AP_rare	mAP_overall, final	Label efficiency
LVIS Random	12.6	14.6 ± 0.32	+2.0	34.1	33%
LVIS Uncertainty	12.6	17.6 ± 0.25	+5.0	35.4	33%
LVIS TAAST	12.6	18.9 ± 0.20	+6.3	36.0	42%
nuImages Random	21.3	23.2 ± 0.34	+1.9	41.6	34%
nuImages Uncertainty	21.3	26.3 ± 0.29	+5.0	42.8	34%
nuImages TAAST	21.3	27.7 ± 0.25	+6.4	43.2	43%

Source: compiled by the author based on the results of the experiment

The results show a clear advantage of the TAAST strategy in all aspects considered. The increase in the AP_rare metric on both datasets was more than + 6 percentage points, exceeding the "net" uncertainty by about + 1.3 percentage points. This confirms the effectiveness of combining entropy selection with logarithmically weighted selection of rare classes. The overall mAP has also grown, which means that there is no degradation in common classes. In addition, Label Efficiency exceeded 42-43%, which indicates a significant reduction in the need for manual markup. Since the 95% confidence intervals between TAAST and Uncertainty-only do not overlap, the difference is statistically significant.

The results obtained confirmed the effectiveness of the TAAST strategy in the context of long-tail active self-learning tasks. Compared to classic uncertainty-based selection scenarios (Settles, 2009), TAAST provides a significantly higher increase in accuracy in rare categories (AP_rare), while maintaining or even improving the overall mAP. This indicates an effective integration of pseudo-labels and frequency weighting and an analysed reduction in the need for manual marking. The current results are consistent with the findings by K. Sohn *et al*. (2020), which showed that high-quality pseudomarking combined with self-learning can be effective, although they did not consider the class imbalance. TAAST extends this idea by adding adaptive weighting over the frequency of classification categories.

In addition, the results are consistent with a number of new approaches in long-tail detection and active learning. In particular, the Plug-and-Play Active Learning (PPAL) method (Yang *et al*., 2024) implements a two-step sampling scheme focused on sample diversity, which is easily integrated into standard detection pipelines without significant architectural changes and provides stable AP growth with minimal overhead. In the field of 3D detection, the Rare Example Mining (REM) approach, proposed by C.M. Jiang *et al*. (2022), addresses the intra-class long tail by purposefully selecting rare examples: the combination of data-centric and model-centric steps allows achieving performance close to fully marked models, with significantly less manual labels. The long-tail problem in unmanned driving is systematically formalised by the LT3D method presented by N. Peri *et al*. (2023): hierarchical loss and multimodal RGB + LiDAR Fusion have been shown to significantly improve the accuracy of rare classes (such as "stroller") by better distinguishing small objects. Ultimately, in the broader context of the open long tail of OLTR++, Z. Liu *et al*. (2022) proposed an integrated framework with dynamic meta-embedding and modular active learning that simultaneously covers the imbalance, few-shot, and open-set aspects — the results were confirmed on large ImageNet, Places, and MS1M sets.

Thus, the experimental results demonstrate that the TAAST strategy can combine the benefits of active and self-learning approaches in a single cycle. It allows significantly reducing the number of frames that require manual annotation, while not losing the overall accuracy of the model. It is important to note that the integration of frequency weighting directly affects the balance of the distribution of selected examples, which is crucial for improving rare classes. This suggests that not only architectural solutions, but also the process of forming a training set itself can become a key factor in improving the efficiency of object detection systems. Overall, the study not only confirmed the effectiveness of active learning as a concept, but also showed that the consideration of class frequency and semantic diversity allows for a much better balance between model quality and annotation costs. Thus, the proposed approach solves one of the key problems of active learning – the preference for frequent classes in the selection process – and offers a practical solution for problems with an imbalanced distribution that often occur in real-world conditions.

Conclusions

The proposed Tail-Aware Active Self-Training method confirmed the effectiveness of targeted and informative sampling in active self-learning tasks. Unlike classical strategies that focus only on entropy or randomness of choice, TAAST combines a rarity weighting factor in combination with an entropy estimation of model uncertainty, which allows prioritising frames with underrepresented classes. This approach provided an increase in average accuracy for rare objects by 6.3-6.4 percentage points, exceeding the results of both random and entropy active learning strategies. Using the 0.8 threshold for pseudo-marking helped to automatically include up to 50% of objects in the training set without the need for manual marking, which resulted in human resource savings of up to 43%. However, the quality of the model did not deteriorate, but on the contrary – it increased both at the level of rare classes and at the level of the overall average indicator. A key role in achieving high efficiency was played by the use of the CLIP model, which allows evaluating the semantic similarity of images without additional training, and clustering by the *k*-means method, which provided grouping scenes in seconds. This allowed avoiding duplication of frames and guarantee maximum diversity in the sample. It was shown that the model reaches a plateau of quality growth after the second iteration of active learning, which allows limiting further additional learning to 30-epoch cycles without losing productivity.

Thus, the method significantly speeds up the detector's self-learning, reduces computational costs, and minimises dependence on expensive manual annotations, making it suitable for use in real-world production environments with a limited budget. Prospects for further research are to extend the method to multimodal data sets, where text or sensory information is available in addition to images. It is also advisable to explore the possibility of adapting TAAST to video streams, where the time context can improve the quality of frame selection. In addition, an important area is the development of an automated mechanism for dynamically adjusting weight coefficients depending on changes in the statistics of marked-up data during the active cycle.

Acknowledgements

None.

Funding

The study received no funding.

Conflict of Interest

None.

References

[1] Ali, M.L., & Zhang, Z. (2024). The YOLO framework: A comprehensive review of evolution, applications, and benchmarks in object detection. *Computers*, 13(12), article number 336. doi: 10.3390/computers13120336.

[2] Bottou, L. (2012). Stochastic gradient descent tricks. In G. Montavon, G.B. Orr & K.R. Müller (Eds.), *Neural networks: Tricks of the trade*. *Lecture notes in computer science* (Vol. 7700, pp 421-436). Berlin: Springer. doi: 10.1007/978-3-642-35289-8_25.

[3] Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., & Beijbom, O. (2020). nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 11621-11631). Seattle: IEEE/CVF. doi: 10.1109/CVPR42600.2020.01164.

[4] De Alvis, C., & Seneviratne, S. (2024). A survey of deep long-tail classification advancements. *ArXiv*. doi: 10.48550/arXiv.2404.15593.

[5] Duan, C.-L., Li, Y., Wei, X.-S., & Zhao, L. (2024). Longtail object detection pre-training: Dynamic rebalancing contrastive learning with dual reconstruction. In *38th conference on neural information processing systems (NeurIPS 2024)*. Vancouver: NeurIPS.

[6] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of Machine Learning Research*, 48, 1050-1059.

[7] Jocher, G., Chaurasia, A., & Qiu, J. (2023). *YOLOv5 and YOLOv8: A detailed comparison*. Retrieved from https://docs.ultralytics.com/models/yolov8/.

[8] Johnson, J., Douze, M., & Jégou, H. (2021). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535-547. doi: 10.1109/TBDATA.2019.2921572.

[9] Li, B., Yao, Y., Tan, J., Zhang, G., Yu, F., Lu, J., & Luo, Y. (2022). Improving long-tailed object detection with image-level supervision by multi-task collaborative learning. *ArXiv*. doi: 10.48550/arXiv.2210.05568.

[10] Li, Y., Wang, T., Kang, B., Tang, S., Wang, Ch., Li, J., & Feng, J. (2020). Overcoming classifier imbalance for long-tail object detection via balanced group softmax. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10991-11000). Seattle: IEEE. doi: 10.1109/CVPR42600.2020.01100.

[11] Qi, T., Xie, H., Li, P., Ge, J., & Zhang, Y. (2023). Balanced classification: A unified framework for long-tailed object detection. *IEEE Transactions on Multimedia*, 26, 3088-3101. doi: 10.1109/TMM.2023.3306968.

[12] Radford, A., *et al*. (2021). Learning transferable visual models from natural language supervision. In *38th international conference on machine learning (ICML 2021)* (pp. 8748-8763). Online Conference.

[13] Sener, O., & Savarese, S. (2018). Active learning for convolutional neural networks: A core-set approach. In *ICLR 2018 conference track: 6th international conference on learning representation*. Vancouver: Vancouver Convention Center.

[14] Settles, B. (2009). *Active learning literature survey*. Madison: University of Wisconsin-Madison.
[15] Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., & Raffel, C. (2020). FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *NIPS'20: Proceedings of the 34th international conference on neural information processing systems* (pp. 596-608). Vancouver: NIPS.
[16] Tian, Z., Shen, C., Chen, H., & He, T. (2019). FCOS: Fully convolutional one-stage object detection. In *IEEE/CVF international conference on computer vision (ICCV)* (pp. 9626-9635). Seoul: IEEE. doi: 10.1109/ICCV.2019.00972.
[17] Wu, J., Chen, J., & Huang, D. (2022). Entropy-based active learning for object detection with progressive diversity constraint. In *IEEE/CVF conference on computer vision and pattern recognition* (CVPR) (pp. 9387-9396). New Orleans: IEEE. doi: 10.1109/CVPR52688.2022.00918.
[18] Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., & Liu, Z. (2021). End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 3060-3069). Montreal: IEEE. doi: 10.1109/ICCV48922.2021.00305.
[19] Yang, C., Huang, L., & Crowley, E.J. (2024). Plug-and-play active learning for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2024)* (pp. 17784-17793). Seattle: IEEE. doi: 10.1109/CVPR52733.2024.01684.
[20] Jiang, C.M., Najibi, M., Qi, C.R., Zhou, Y., & Anguelov, D. (2022). Improving the intra-class long-tail in 3D detection via rare example mining. In *Computer vision – ECCV 2022. Lecture notes in computer science* (Vol. 13670, pp. 155-172). Cham: Springer. doi: 10.1007/978-3-031-20080-9_10.
[21] Peri, N., Dave, A., Ramanan, D., & Kong, S. (2023). Towards long-tailed 3d detection. *Proceedings of Machine Learning Research,* 205, 1904-1915.
[22] Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., & Yu, S.X. (2022). Open long-tailed recognition in a dynamic world. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 46(3), 1836-1851. doi: 10.1109/TPAMI.2022.3200091.

Активне самонавчання для детекції об'єктів в умовах дисбалансованих даних: підхід TAAST

Дмитро Іванов
Аспірант
Державний університет «Житомирська політехніка»
10005, вул. Чуднівська, 103, м. Житомир, Україна
https://orcid.org/0000-0002-7386-4497

Анотація. У контексті дедалі ширшого розвитку та застосування комп'ютерного зору зростає потреба у зменшенні витрат на ручну розмітку даних, особливо в задачах виявлення рідкісних об'єктів за умов довгохвостого розподілу класів. Метою дослідження було підвищення ефективності визначення рідкісних категорій зображень через вдосконалення стратегії активного самонавчання. У роботі застосовано підхід Tail-Aware Active Self-Training, що базується на стратегічному відборі кадрів з урахуванням ентропії невпевненості, рідкісності класу та семантичного різноманіття в просторі ознак моделі Contrastive Language-Image Pretraining, з подальшим використанням псевдорозмітки за допомогою детектора You Only Look Once, версія 8. У результаті експериментів на наборах даних Large Vocabulary Instance Segmentation, версія 1.0 та nuImages-imbalanced запропонована стратегія забезпечила приріст точності AP_rare на 6,3–6,4 відсоткових пунктів у порівнянні з базовими підходами Random та Uncertainty Sampling. Загальна точність моделі при цьому не знизилась, а зросла до 36,0–43,2 % mAP залежно від датасету. Показник ефективності розмітки досягнув 42–43 %, що на 9-10 пунктів вище за конкурентні стратегії. Результати експерименту є статистично достовірними, оскільки інтервали довіри для метрики точності AP_rare у разі застосування методу Tail-Aware Active Self-Training не перетинаються з інтервалами для базових стратегій Random і Uncertainty-only. Це свідчить про те, що перевага даного методу не є випадковою, а підтверджена з високою ймовірністю. Отже, отримані результати продемонстрували надійність і стабільність запропонованого підходу: вже після двох активних ітерацій модель досягла плато продуктивності, що дозволило суттєво зменшити обчислювальні витрати. Практична цінність роботи полягає у створенні ефективного інструменту для автоматизованого розгортання моделей комп'ютерного зору в умовах обмеженого бюджету на розмітку

Ключові слова: машинне навчання; семантична кластеризація; псевдоанотація; вибірка за ентропією; балансування класів; комп'ютерний зір; оптимізація розмітки