NegRefine: Refining Negative Label-Based Zero-Shot OOD Detection

Amirhossein Ansari[1] Ke Wang[1] Pulei Xiong[2]
[1]Simon Fraser University [2]National Research Council Canada
ah_ansari@sfu.ca, wangk@cs.sfu.ca, pulei.xiong@nrc-cnrc.gc.ca

Abstract

Recent advancements in Vision-Language Models like CLIP have enabled zero-shot OOD detection by leveraging both image and textual label information. Among these, negative label-based methods such as NegLabel and CSP have shown promising results by utilizing a lexicon of words to define negative labels for distinguishing OOD samples. However, these methods suffer from detecting in-distribution samples as OOD due to negative labels that are subcategories of in-distribution labels or proper nouns. They also face limitations in handling images that match multiple in-distribution and negative labels. We propose NegRefine, a novel negative label refinement framework for zero-shot OOD detection. By introducing a filtering mechanism to exclude subcategory labels and proper nouns from the negative label set and incorporating a multi-matching-aware scoring function that dynamically adjusts the contributions of multiple labels matching an image, NegRefine ensures a more robust separation between in-distribution and OOD samples. We evaluate NegRefine on large-scale benchmarks, including ImageNet-1K. The code is available at https://github.com/ah-ansari/NegRefine.

1. Introduction

Machine learning models deployed in real-world environments often encounter input samples from unknown classes that differ significantly from the pre-defined classes in the training data. Model predictions for these Out-of-Distribution (OOD) samples are typically unreliable and overconfident, posing critical risks and security vulnerabilities in critical domains like healthcare and autonomous driving [27]. OOD detection mitigates this issue by distinguishing OOD samples from in-distribution ones, enabling models to reject and flag OOD inputs when encountered [1, 12, 41]. Traditionally, image OOD detection methods have relied on single-visual modality [9, 20, 21, 39], neglecting the textual information available in class labels. However, with recent advancements in pre-trained Vision-Language Models like CLIP [30], adopting a multi-modal paradigm that also incorporates class label information has gained significant attention [10, 22, 38]. Among these, negative label-based methods, such as NegLabel [16] and CSP [6], have emerged as particularly promising solutions.

The core idea behind NegLabel [16] is to leverage a large lexicon of English words, such as WordNet [11], as a label pool to create a set of negative (OOD) labels Y_{neg}. To form Y_{neg}, words (nouns and adjectives) from the lexicon are ranked based on their semantic similarity to the in-distribution labels in Y_{in}, and the top $p\%$ least similar words to the in-distribution labels are selected as negative labels. For an input image, the similarity between the image and both the in-distribution labels in Y_{in} and the negative labels in Y_{neg} is computed using CLIP, and based on these similarity scores, a final score is defined to detect the image as in-distribution or OOD. CSP [6] improves upon NegLabel by refining the use of adjectives to construct superclass labels that better capture OOD samples.

Despite the impressive performance of NegLabel [16] and CSP [6], these methods still face the following notable limitations:

Subcategory Overlap. These methods rely on textual similarity and a cutoff of the top $p\%$ least similar words to select Y_{neg}. This approach is inherently limited, as it depends solely on semantic similarity without explicitly considering hierarchical relationships among words. As a result, they can mistakenly include subcategories (*e.g.*, "blue-eyed african daisy") of in-distribution labels (*e.g.*, "daisy") in the negative set, leading to an overlap between Y_{neg} and Y_{in}. This issue is particularly pronounced because, for a given image, CLIP tends to assign higher similarity scores to more specific labels than to general terms, consequently, the negative subcategory label can receive a higher similarity score than the more general in-distribution label, leading to incorrect detection of in-distribution samples as OOD. Fig. 1a shows this issue using an in-distribution sample from the ImageNet-1K [8] dataset.

Proper Nouns. WordNet contains many proper nouns (*e.g.*, names of individual persons, places, or organizations), leading to their excessive selection as negative labels. The CLIP model tends to match samples from cer-

(a) Subcategory Overlap (b) Proper Noun (c) Multi-Label Matching (d) Multi-Label Matching

Figure 1. In-distribution images from ImageNet-1K with the top two labels from Y_{in} and Y_{neg}. The value beside each label represents CLIP's similarity score multiplied by 100. These images are incorrectly detected as "OOD" by CSP due to: (a) a subcategory of the in-distribution label appearing in the negative labels, (b) proper nouns in the negative labels receiving high similarity scores. (c) and (d) multi-label matching issue, with negative matching labels receiving higher similarity scores.

tain in-distribution classes with proper nouns in Y_{neg} with very high similarity, resulting in their incorrect detection as OOD. Fig. 1b shows an image from the in-distribution class "beach", for which CLIP assigns a much higher similarity score to some proper nouns in Y_{neg}, such as "costa rica" (a country) and "tobago" (an island), than to the true label "beach" from Y_{in}, even though the image is not actually from "costa rica" or "tobago". One likely reason is that these locations are well known for their beaches or strongly associated with beach images in CLIP's training data.

Multi-Label Matching. A real-world image often contains multiple objects [25], matching the labels of all the objects in the image. Even a single-object image can match multiple labels, with each label describing a different aspect of the image. In such cases, if an image from an in-distribution class also matches some negative labels, it may be incorrectly detected as OOD by NegLabel and CSP. This is because these methods rely solely on the significance of CLIP's similarity scores for individual labels in Y_{in} and Y_{neg}, without accounting for multiple labels matching an image. Figs. 1c and 1d illustrate this issue using two in-distribution images: (c) a multi-object image containing both a "bee" (in-distribution label) and a flower "tickseed sunflower" (negative label), and (d) a single-object image of a "necklace" (in-distribution label) with the shape of a leaf, matching "pinnate-leaved item" (negative label). These images are incorrectly detected as OOD by CSP because the negative labels received higher similarity scores.

In this paper, we propose NegRefine, a novel framework that addresses the limitations of existing negative label-based methods for zero-shot OOD detection. Our contributions are as follows:

- NegRefine incorporates a filtering mechanism to prevent the inclusion of subcategory labels and proper nouns in the negative label set, ensuring a clearer semantic separation between Y_{in} and Y_{neg}.
- NegRefine introduces a novel scoring function that accounts for multiple labels matching an image, dynamically adjusting their contributions based on contextual relevance to mitigate the misprediction of in-distribution images as OOD when they also match negative labels.
- We extensively evaluate NegRefine on large-scale benchmarks, including ImageNet-1K. As noted in [3], a significant portion of existing OOD datasets belongs to one of the classes in ImageNet-1K, making them unsuitable as true OOD samples. Following [3], we adopt more reliable OOD datasets for ImageNet-1K. Our method outperforms the state-of-the-art CSP [6] by 1.82% increase in AUROC and 4.35% reduction in FPR95.

2. Preliminaries

2.1. Problem Definition

Given in-distribution class labels $Y_{in} = \{y_1, y_2, \ldots, y_K\}$, where y_i is the textual label for class i and K is the number of classes, CLIP-based zero-shot OOD detection [10, 16, 22, 38] is to detect whether an input image x is "in-distribution" (*i.e.*, belongs to one of the classes in Y_{in}) or "OOD", using the CLIP model and Y_{in} as input. In the zero-shot setting, no training data is available to fine-tune CLIP or adjust its text prompt. Solving this problem requires designing an in-distribution scoring function $S(x)$, which assigns higher scores to in-distribution samples and lower scores to OOD samples. Given a threshold γ, an image x is detected as in-distribution if $S(x) \geq \gamma$, and as OOD otherwise.

2.2. CLIP Model

CLIP [30] is a vision-language model trained on a large corpus of (image, caption) pairs using a contrastive loss to align images with their corresponding text representations in a shared embedding space. It consists of an image encoder, $f_{\text{img}} : x \to \mathbb{R}^D$, and a text encoder, $f_{\text{txt}} : t \to \mathbb{R}^D$, which map an image x and a text t to a D-dimensional feature space. In standard CLIP-based zero-shot classification, a text prompt is first created for each class label $y_i \in Y$ in a format like "This is a y_i" (*e.g.*, "This is a dog"). The cosine similarity is then computed between the input image embedding, $f_{\text{img}}(x)$, and each text prompt embedding, f_{txt}("This is a y_i"). The image x is assigned to the class label with the highest cosine similarity.

2.3. Negative Label-Based OOD Detection

NegLabel [16] introduced creating a set of negative (OOD) labels, Y_{neg}, to fully leverage CLIP for zero-shot OOD detection. To select the negative labels, nouns and adjectives from WordNet were ranked based on their similarity to the in-distribution labels, computed using CLIP's text encoder. The top $p = 15\%$ least similar words to the in-distribution labels were then selected as the negative label set Y_{neg}. Using Y_{in} and Y_{neg}, they defined their in-distribution score as:

$$S_{NegLabel}(x) = \frac{\sum_{i=1}^{K} e^{\text{sim}(x,y_i)/\tau}}{\sum_{i=1}^{K} e^{\text{sim}(x,y_i)/\tau} + \sum_{j=1}^{M} e^{\text{sim}(x,\tilde{y}_j)/\tau}}, \tag{1}$$

where y_i represents the i-th in-distribution label (from K in-distribution labels), and $\tilde{y}_j$ is the j-th negative label (from M negative labels). The function $\text{sim}(x, y_i)$ denotes the cosine similarity between the embedding of the image x and the embedding of the text for label y_i. The temperature parameter is set to $\tau = 0.01$ to scale the similarity scores. CSP [6] follows the same structure as NegLabel but enhances it by leveraging adjectives from WordNet to generate modified superclass labels, which better match OOD samples.

3. Methodology

We propose NegRefine to address the limitations of current negative label-based methods discussed in the Introduction, *i.e.*, the inclusion of subcategories of in-distribution labels and proper nouns in the negative set, and the inability to effectively handle images matching multiple labels. NegRefine has two key components: (i) a negative label filtering mechanism to improve the quality of negative labels, and (ii) an effective multi-matching score that dynamically adjusts to handle in-distribution images matching multiple labels. The outcome is a new score function S based on the refined negative label set produced by (i) and the adjustment applied by (ii). With this new score, the OOD detection is the same as described in Sec. 2.1. In the rest of this section, we discuss the negative label filtering mechanism and the multi-label matching score function.

3.1. Negative Label Filtering Mechanism

As discussed, subcategory labels like "african daisy" receive higher similarity than their more general in-distribution labels (Fig. 1a), and proper nouns like "costa rica" can match certain in-distribution images (Fig. 1b), leading to incorrect detection of such images as OOD. To mitigate these issues, we propose to remove such negative labels. Note that removing proper nouns does not impact OOD detection, as proper nouns are so specific that if an OOD image is truly related to a proper noun, a more general non-proper noun negative label (*e.g.*, "island" instead of "tobago") will likely capture the image. This is because a good negative set is expected to cover all OOD images.

Our negative label filtering method, NegFilter, is presented in Algorithm 1. It starts with the initial set of negative labels selected in CSP [6], denoted as Y_{neg}, and refines it by filtering out improper negative labels, *i.e.*, proper nouns and subcategories. To identify such labels, we leverage an off-the-shelf Large Language Model (LLM) to automate the filtering process. LLMs have demonstrated near-human performance in various NLP tasks [5, 26], making them a reliable tool for this purpose. For each term w in Y_{neg}, we apply the following two filtering checks:

- **Proper Nouns Check:** We query the LLM with the prompt "Is w a proper noun, like the name of an entity?" (line 3). If the response is "Yes", we remove w from Y'_{neg} (line 5).
- **Subcategory Check:** We first identify the set L of the top n most similar labels in Y_{in} to w using CLIP's text encoder (lines 7–12) to minimize the number of subcategory queries. Then, for each $l \in L$, we query the LLM with the prompt: "Is w a subcategory of l?". If the LLM confirms that w is a subcategory of any label in this set, we remove w from Y'_{neg} (line 11).

The resulting negative label set Y'_{neg} has lower similarity to in-distribution images, reducing the likelihood of misdetecting in-distribution samples as OOD.

WordNet [11] also provides a hierarchy of words that can be used to identify subcategories. However, we observed several issues that make WordNet unreliable for our task: "african daisy" and "daisy" are at the same level instead of forming a parent-child pair; closely related words are placed in different branches, such as some subcategories of "mushroom" fall under "fungus", while others fall under "fungus-genus" (fungus is the general term for mushrooms). Moreover, WordNet cannot be used to identify proper nouns. Therefore, we opted to use LLMs for their higher accuracy, greater flexibility, and ease of use.

Algorithm 1 NegFilter($Y_{neg}, Y_{in}, LLM, f_{\text{txt}}, n$)

Input: Y_{neg} (initial negative label set), Y_{in} (in-distribution label set), LLM (Large Language Model), f_{txt} (CLIP text encoder), n (number of in-distribution labels for subcategory check).

Output: Y'_{neg} (refined negative label set).

$Y'_{neg} \leftarrow Y_{neg}$
for each $w \in Y_{neg}$ **do**
  $R \leftarrow LLM($ "Is w a proper noun, like the name of an entity?")
  if $R =$ "Yes" **then**
    $Y'_{neg} \leftarrow Y'_{neg} - \{w\}$
  else
    $L \leftarrow n$ top similar labels to w in Y_{in} using f_{txt}
    for each $l \in L$ **do**
      $R \leftarrow LLM($ "Is w a subcategory of l?")
      if $R =$ "Yes" **then**
        $Y'_{neg} \leftarrow Y'_{neg} - \{w\}$
        break
return Y'_{neg}

3.2. Multi-Label Matching Score Function

The in-distribution score used in NegLabel and CSP (Eq. (1)) relies on the similarity scores of individual in-distribution and negative labels, but it overlooks cases where in-distribution images also match negative labels with high similarity (see Figs. 1c and 1d). To address this issue, we propose a new score that can adjust and capture such multi-label matching cases.

Consider the training paradigm of the CLIP model. CLIP is trained on (image, caption) pairs, where captions often describe multiple objects or aspects of an image. This training enables CLIP to develop a deep understanding of an image's visual content. Motivated by this, instead of relying on individual label similarity as in CSP and NegLabel, we propose creating a more caption-like text by concatenating an in-distribution label y with a negative label $\tilde{y}$, forming the text "y and $\tilde{y}$". For example, for the image in Fig. 1c, the concatenated text is "bee and tickseed sunflower". If the similarity score for this text significantly increases compared to the score for the negative label "tickseed sunflower" alone, it suggests that the joint description better captures the image content and reinforces the relevance of the in-distribution label "bee". Conversely, if the similarity remains unchanged or decreases, it suggests that the labels do not jointly describe the image, and so the in-distribution label does not contribute.

To illustrate, consider the images in Figs. 1c and 1d, which truly match both in-distribution and negative labels:

- Fig. 1c: the similarity for "bee" (in-distribution) is 30.2, and for "tickseed sunflower" (negative) is 31.9, while the similarity increases for "bee and tickseed sunflower" to 34.4.
- Fig. 1d: the similarity for "necklace" (in-distribution) is 28.9, and for "pinnate-leaved item" (negative) is 29.1, while the similarity increases for "necklace and pinnate-leaved item" to 34.0.

These examples show how similarity significantly increases for the concatenated text when the two labels truly correspond to the image.

To formalize our approach, let T_{in} and T_{neg} denote the top k in-distribution and negative labels that best match an image based on CLIP similarity scores. For each $y_i \in T_{in}$ and $\tilde{y}_j \in T_{neg}$, we create the text $t_{i,j}$ as "y_i and $\tilde{y}_j$". With k in-distribution and k negative labels, this approach generates k^2 text combinations. We define a multi-matching score S_{MM} to measure the contribution of the in-distribution label y_i through the comparison between the concatenated text and the negative label alone. For an input image x, $S_{MM}(x)$ is defined as:

$$S_{MM}(x) = \max_{i,j} \frac{e^{\text{sim}(x,t_{i,j})/\tau}}{e^{\text{sim}(x,t_{i,j})/\tau} + e^{\text{sim}(x,\tilde{y}_j)/\tau}}, \quad (2)$$

where $\text{sim}(x, t_{i,j})$ denotes the cosine similarity between the embeddings of the image x and text $t_{i,j}$.

For a multi-matching in-distribution image, the in-distribution label y_i is truly relevant to the image, thus, the similarity of $t_{i,j}$ ("y_i and $\tilde{y}_j$") will be higher than that of $\tilde{y}_j$ alone, resulting in a larger S_{MM} score. Conversely, for an OOD image, the in-distribution label y_i is not relevant, so the similarity of $t_{i,j}$ will be equal to or lower than that of $\tilde{y}_j$, leading to a smaller S_{MM} score.

Final Score. The initial in-distribution score from NegLabel, $S_{NegLabel}$ (see Eq. (1)), applies a softmax and summation over individual in-distribution and negative label similarity scores to measure the overall attraction toward in-distribution or OOD. In contrast, we propose S_{MM} as an additional adjustment to capture multi-matching cases. We combine the collective summation from $S_{NegLabel}$ with the multi-matching adjustment from S_{MM} to define our final in-distribution score as:

$$S(x) = S_{NegLabel}(x) + \alpha \times S_{MM}(x). \quad (3)$$

Here, α is a weighting factor that controls the contribution of the multi-matching score S_{MM} to the final score.

4. Experiments

Datasets and Benchmarks. The main evaluation in most previous studies [6, 16, 38] has been conducted using the ImageNet-1K [8] dataset as the in-distribution data, with subsets of iNaturalist [34], SUN [40], Places [44], and the full Textures [7] dataset, used as OOD [15]. However, [3] randomly selected 400 samples from 12 commonly used

Methods	OOD Datasets									
	iNaturalist		OpenImage-O		Clean		NINCO		Average	
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
ZOC [10]	79.53	90.98	80.61	70.04	72.88	81.07	67.96	78.67	75.24	80.19
MCM [22]	94.59	32.20	92.00	41.03	83.24	57.79	74.34	79.33	86.04	52.59
GL-MCM [25]	96.44	17.42	92.91	34.52	84.78	51.49	76.03	74.40	87.54	44.46
CLIPN [38]	96.20	19.11	92.22	30.36	87.31	41.56	78.72	66.51	88.61	39.39
NegLabel [16]	99.49	1.91	93.74	28.62	86.79	41.22	77.30	68.70	89.33	35.11
CSP [6]	**99.60**	1.54	94.09	28.94	88.32	38.75	77.88	68.65	89.97	34.47
NegRefine (ours)	99.57	**1.47**	**95.00**	**23.85**	**90.65**	**33.18**	**81.92**	**61.96**	**91.79**	**30.12**

Table 1. OOD detection performance on the ImageNet-1K in-distribution benchmark. The average is over the four OOD datasets. All methods are zero-shot and utilize CLIP ViT-B/16 model. The metrics are reported as percentages, with the best results highlighted in bold.

OOD datasets and, through manual examination, found that 59.5% of the samples from Places and 25.6% from Textures were actually in-distribution (*i.e.*, they could be assigned to one of the classes in ImageNet-1K). Among all 12 datasets analyzed, only iNaturalist [34] (2.5%) and OpenImage-O [37] (4.9%) exhibited significantly lower in-distribution rates. We extended this analysis to the SUN dataset, which was not covered in [3], and observed an in-distribution rate of 26.2%. To address this high in-distribution rate, the same study [3] introduced the NINCO dataset as a more reliable OOD data for ImageNet-1K and also provided the set of clean OOD samples obtained from their manual analysis of 400 samples across the 12 datasets, which we refer to as Clean. In our evaluation, we consider NINCO and Clean and also include iNaturalist and OpenImage-O, which contain minimal in-distribution contamination. Thus, our OOD datasets for the ImageNet-1K benchmark are iNaturalist, OpenImage-O, Clean, and NINCO. A description of these OOD datasets is provided in Supplementary Sec. A.1.

Implementation Details. We employ the CLIP ViT-B/16 [30] as the pre-trained CLIP model. For the initial negative labels, label prompts for CLIP, and other parameter details, we followed the settings of CSP [6]. For our negative label filtering mechanism (Algorithm 1), we use Qwen2.5-14B-Instruct [2] as the LLM and set $n = 10$. For the multi-matching score S_{MM}, we use $k = 5$ and set $\alpha = 2$ as the weight of S_{MM} in Eq. (3). All experiments are conducted on NVIDIA RTX 6000 GPUs.

Evaluation Metrics. Following previous studies [9, 10, 38], we evaluate our method using the Area Under the Receiver Operating Characteristic Curve (AUROC) and the False Positive Rate at the 95% True Positive Rate, denoted by FPR95. In these measures, in-distribution data is treated as the positive class.

Baselines. We compare our method, NegRefine, with state-of-the-art CLIP-based zero-shot OOD detection baselines listed in the first column of Tab. 1. See Related Work for details on these baselines. In this comparison, we exclude few-shot and supervised training methods, as they require training data, whereas we focus on the zero-shot setting.

4.1. Main Results

Tab. 1 presents the results comparing the performance of our method (NegRefine) against previous CLIP-based zero-shot OOD detection studies on the ImageNet-1K benchmark. Our method outperforms the state-of-the-art approach, CSP, achieving a 1.82% increase in average AUROC and a 4.35% reduction in average FPR95. More specifically, out of the eight AUROC and FPR95 metrics evaluated across the four OOD datasets, our method achieves the best result in seven cases, while CSP only slightly outperforms us by 0.03% on AUROC for iNaturalist. Results for our method are averaged over 10 different seeds, with standard deviations reported in Supplementary Tab. 6. For details about the implementation of baselines and the source of randomness in our method and other baselines, refer to Supplementary Secs. A.2 and A.3.

4.2. Analysis

Analysis of Main Components. Our method introduces two key components: (i) NegFilter (Algorithm 1) to remove negative labels that are proper nouns and subcategories of in-distribution labels, and (ii) the S_{MM} score (Eq. (2)) to handle in-distribution images matching multiple labels. Tab. 2 presents an ablation study analyzing the individual contribution of each component. The first row corresponds to the CSP method, serving as the reference for comparison. Applying negative label filtering alone improves the average FPR95 by 2.59%, while incorporating S_{MM} alone reduces FPR95 by 2.77%. Integrating both components in our final method (last row) results in a 4.35% reduction in average FPR95.

Analysis of Individual Filters in NegFilter. In Tab. 3, we analyze the contributions of the two filtering checks used in our negative label filtering (NegFilter): proper nouns filtering, denoted by C1, and subcategory filtering, denoted by C2. S_{MM} is always applied as the base case in this study. C1 alone reduces the average FPR95 by 0.61%, while C2

S_{MM}	NegFilter	iNaturalist		OpenImage-O		Clean		NINCO		Average	
		AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
×	×	99.60	1.54	94.09	28.94	88.32	38.75	77.88	68.65	89.97	34.47
×	✓	**99.66**	**1.28**	94.92	24.93	89.41	35.37	79.83	65.93	90.95	31.88
✓	×	99.46	1.74	94.52	26.23	89.99	35.26	81.42	63.57	91.35	31.70
✓	✓	99.57	1.47	**95.00**	**23.85**	**90.65**	**33.18**	**81.92**	**61.96**	**91.79**	**30.12**

Table 2. Analysis of the two main components of our method. × denotes that the component is turned off, and ✓ indicates it is applied. The first row, *i.e*. all components are off, corresponds to CSP.

NegFilter		iNaturalist		OpenImage-O		Clean		NINCO		Average	
C1	C2	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
×	×	99.46	1.74	94.52	26.23	89.99	35.26	81.42	63.57	91.35	31.70
✓	×	99.51	1.62	94.78	25.15	90.40	34.34	81.63	63.27	91.58	31.09
×	✓	99.54	1.63	94.78	24.81	90.29	34.06	81.65	62.80	91.56	30.82
✓	✓	**99.57**	**1.47**	**95.00**	**23.85**	**90.65**	**33.18**	**81.92**	**61.96**	**91.79**	**30.12**

Table 3. Analysis of the filtering checks used in our negative label filtering component. C1 refers to the proper nouns check, and C2 refers to the subcategory check. S_{MM} is always applied.

reduces it by 0.88%. When applied together, they achieve a combined reduction of 1.58% in average FPR95.

The proper nouns check removes 1749 negative labels, while the subcategory check removes 307 negative labels, accounting for 20.6% and 3.6% of the initial set of negative labels, respectively. Examples of removed proper nouns are costa rica, tobago, golden gate bridge, budapest, and saint christopher, and examples of removed subcategories are african daisy (daisy), morchella (mushroom), calendula (daisy), tandoor (stove), with the corresponding supercategory in-distribution labels shown in parentheses. The presence of these erroneous negative labels can lower the in-distribution score for in-distribution samples, leading to misdetection, as discussed in Figs. 1a and 1b.

Analysis of the Effect of S_{MM}. Fig. 2 illustrates how incorporating S_{MM} improves OOD detection using two in-distribution examples (a) and (b) from ImageNet-1K and an OOD example (c) from the NINCO dataset. Using the original score $S_{NegLabel}$, all three examples are predicted incorrectly due to the multi-label matching issue. By applying S_{MM}, that is, using the new score $S = S_{NegLabel} + 2S_{MM}$, all samples are correctly predicted. The reason for this improvement lies in how S_{MM} behaves for different types of samples. For the in-distribution images (a) and (b), the change in similarity from the negative label alone to the multi-matching text is significant, 6.24 in (a) and 4.79 in (b), resulting in a high S_{MM} score. This large change is because, for these examples, both in-distribution and negative labels match the image, making the combined text a better description for the content of the image (refer to Sec. 3.2). Conversely, for the OOD sample (c), this change is smaller, only 1.58, leading to a low S_{MM} score. In this way, S_{MM} adjusts the final score to correct the prediction.

Comparing S_{MM} to CLIP's Local Features. Recently, GL-MCM [25] proposed using CLIP's local embeddings alongside its global (final) embedding for OOD detection to address images with multiple objects, a motivation similar to our S_{MM} score. In CLIP ViT-B/16, these local embeddings are the 14×14 patch tokens, while the global embedding refers to the final [CLS] token output [45]. Their intuition is that local embeddings better capture local objects, thus improving detection for images with multiple objects. Tab. 1 showed that our method outperforms GL-MCM.

For a more direct comparison with our S_{MM}, we extend their local features idea to the negative label-based score by replacing the global image embedding with local embeddings in Eq. (1). We compute the score for each of the 14×14 local embeddings and take the maximum as the local score $S_L(x)$. Using this, we define a new in-distribution score $S'(x) = S_{NegLabel}(x) + \alpha S_L(x)$. We compare $S'(x)$ with our score $S(x) = S_{NegLabel}(x) + \alpha S_{MM}(x)$ in Fig. 3. For a fair comparison, negative label filtering is applied for both scores. Our S outperforms S' across all α values. This superior performance can be attributed to two main reasons: First, S_{MM} addresses more general multi-label matching cases; for example, even single-object images can match multiple labels (see Fig. 1d), where the local-features approach would be ineffective. Second, our score, which uses the global embedding, aligns more closely with CLIP's training, as the global embedding is explicitly optimized to match the corresponding caption.

Extended Analysis and Ablations. Further analyses and

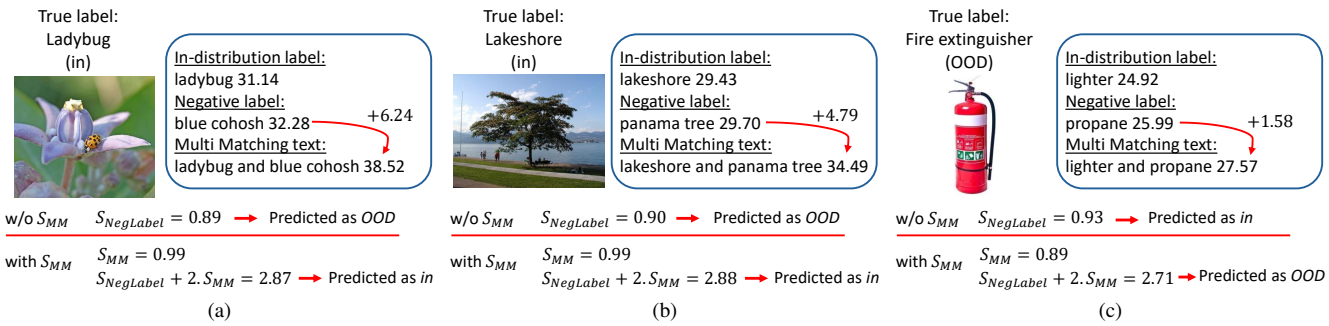

Figure 2. Effect of incorporating S_{MM}. The detection threshold γ (set such that 95% of in-distribution data are detected as in-distribution) is 0.91 for $S_{NegLabel}$ and 2.85 for S. (a) and (b) are in-distribution images from ImageNet-1K, and (c) is an OOD sample from NINCO. The text boxes display the top labels and multi-matching texts with their CLIP similarity scores ($\times 100$). Using $S_{NegLabel}$, all three predictions are incorrect due to multi-matching issue. By incorporating S_{MM}, the final score S is adjusted, leading to correct predictions.

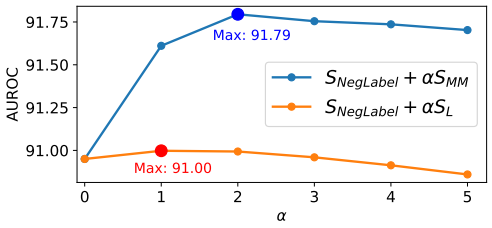

Figure 3. Comparison of our S_{MM} with the local-features-based approach S_L for addressing the multi-matching issue. AUROC ↑, averaged over the OOD datasets across different α, is reported.

ablations on the size of the initial negative label set, the choice of LLM in NegFilter, different CLIP architectures, and alternative designs and parameters for S_{MM} are provided in Supplementary Sec. A.4.

4.3. Additional Benchmarks

ImageNet Subsets and Spurious OOD. Following [6, 16], we also use subsets of ImageNet-1K, namely ImageNet-10, ImageNet-20, and ImageNet-100, created in [22] for evaluation. These subsets contain 10, 20, and 100 classes, respectively. However, we modify ImageNet-100 by removing the "race car" class, resulting in ImageNet-99. This modification is because we observed that "race car" includes samples identical to those in "sports car" from ImageNet-10, creating conflicts when evaluating ImageNet-10 (in) against ImageNet-100 (OOD). Additionally, we evaluate our method on the Spurious OOD benchmark [23], where WaterBirds [32] (in) is tested against Placesbg (OOD). Results for these benchmarks are presented in Tab. 4.

In this study, when using ImageNet-10 as in-distribution data, our performance is slightly lower than CSP and NegLabel. This is because, with only 10 classes, this subset has high separability between in-distribution and OOD data, making the challenges we address less significant and limiting the potential for further improvement. However, our method outperforms CSP and NegLabel in all other cases, particularly in the more challenging scenario of ImageNet-99 (in) vs. ImageNet-10 (OOD), where FPR95 is largely higher for all methods.

Robustness to Domain Shift. In this benchmark, we use domain-shifted versions of ImageNet as the in-distribution data, including ImageNet-Sketch [36], ImageNetV2 [31], ImageNet-A [14], and ImageNet-R [13]. The first two datasets share the same class labels as ImageNet-1K, while the class labels in the other two are subsets of ImageNet-1K. Since our method relies solely on in-distribution class labels, we use the same set of negative labels as for ImageNet-1K in the first two datasets. For the other datasets, we re-select the negative label set based on their in-distribution labels. The results are presented in Tab. 5. In this experiment, our method outperformed CSP and NegLabel on most datasets, except for ImageNet-R, where NegLabel performed slightly better than our method.

Other In-Distribution Datasets. Evaluations using other datasets as in-distribution data, including Stanford-Cars [17], CUB-200 [35], Oxford-Pet [29], and Food-101 [4], are provided in Supplementary Sec. A.5.

5. Related Work

In this section, we review previous studies that utilized CLIP for OOD detection.

Zero-Shot Methods. ZOC [10] develops a caption generator to suggest unseen labels. It then uses CLIP's similarity for in-distribution and unseen labels to define a detection score. MCM [22] applies softmax over CLIP similarity for in-distribution labels and uses the maximum softmax probability as the in-distribution score. CLIPN [38] introduces a "No" text encoder to assist CLIP in understanding prompts like "not containing y_i", and uses this for OOD detection. GL-MCM [25] improves MCM by leveraging CLIP's local features to address images containing multiple objects. This

in-distribution data	ImageNet-10		ImageNet-10		ImageNet-20		ImageNet-99		WaterBirds	
OOD data	ImageNet-20		ImageNet-99		ImageNet-10		ImageNet-10		Placesbg	
Methods	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
NegLabel	98.80	5.00	99.45	1.89	98.04	11.60	89.53	52.89	87.99	29.16
CSP	**99.02**	**3.30**	**99.50**	**1.78**	98.79	3.40	91.21	44.35	92.88	12.07
NegRefine (ours)	98.95	3.89	99.21	2.24	**99.19**	**2.28**	**91.68**	**40.99**	**93.13**	**11.96**

Table 4. OOD detection performance on ImageNet subsets and the Spurious OOD (WaterBirds vs. Placesbg).

in-distribution data	Methods	iNaturalist		OpenImage-O		Clean		NINCO		Average	
		AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
ImageNet-Sketch	NegLabel	99.34	2.24	93.04	30.46	85.42	43.35	74.78	70.30	88.15	36.59
	CSP	**99.49**	**1.60**	**93.74**	30.86	87.68	40.96	76.91	69.87	89.45	35.82
	NegRefine	99.25	1.95	93.60	**29.27**	**89.05**	**37.42**	**79.90**	**66.52**	**90.45**	**33.79**
ImageNetV2	NegLabel	99.40	2.47	92.54	31.86	84.92	44.57	74.01	71.37	87.72	37.57
	CSP	**99.54**	**1.76**	93.06	31.96	86.80	41.55	74.57	71.13	88.49	36.60
	NegRefine	99.51	1.77	**94.25**	**26.38**	**89.57**	**34.95**	**79.96**	**64.36**	**90.82**	**31.87**
ImageNet-A	NegLabel	98.80	4.09	86.56	42.33	82.78	50.61	66.31	79.33	83.61	44.09
	CSP	**99.15**	**2.91**	88.33	43.62	86.36	46.48	70.69	78.21	86.13	42.80
	NegRefine	98.86	3.45	**89.17**	**40.40**	**87.75**	**42.98**	**74.48**	**74.02**	**87.56**	**40.21**
ImageNet-R	NegLabel	99.58	1.60	**97.02**	**15.22**	94.75	**21.36**	**88.74**	**50.88**	**95.02**	**22.27**
	CSP	**99.79**	**0.89**	96.47	18.15	94.17	22.80	85.83	54.59	94.06	24.11
	NegRefine	99.68	0.98	96.75	16.59	**94.77**	21.47	87.08	51.85	94.57	22.72

Table 5. OOD detection performance on domain-shifted versions of ImageNet.

is a motivation similar to that of our S_{MM}. However, the multi-matching issue we address is more general than the presence of multiple objects in an image, as even single-object images can match multiple labels (see Fig. 1d).

NegLabel [16] introduces negative label-based OOD detection by mining negative labels from WordNet [11] and defining an in-distribution score based on CLIP's similarity to in-distribution and negative labels. CSP [6] improves upon NegLabel by introducing a conjugated semantic pool, where adjectives from WordNet are combined with a predefined set of general labels (*e.g.*, area, creature, item) to generate modified superclass labels. These labels can more effectively capture OOD samples. We build on these two studies by addressing their limitations, primarily to reduce the misdetection of in-distribution samples as OOD.

LAPT [43] and AdaNeg [42] build on NegLabel by applying additional techniques: LAPT uses prompt tuning to optimize NegLabel's text prompts, while AdaNeg employs a test-time adaptation strategy. Our work is orthogonal to these two studies, as we improve the internal mechanism of NegLabel. Notably, LAPT and AdaNeg can also be applied on top of our work by replacing NegLabel with our NegRefine to further enhance performance.

Few-Shot and Full-Data Methods. These methods rely on training data (to a different extent) and are therefore not directly related to ours, as we focus on the zero-shot setting. NPOS [33] fine-tunes CLIP's image encoder and generates OOD samples in the feature space to refine the in-distribution/OOD boundary. LoCoOp [24] optimizes prompts by leveraging irrelevant parts of the in-distribution data as OOD signals to enhance in-distribution/OOD separation. NegPrompt [18] learns negative prompts with semantics opposite to in-distribution classes and then performs OOD detection based on similarity to the learned negative prompts. LSN [28] learns class-specific negative prompts that capture features absent in the in-distribution classes. SeTAR [19] enhances OOD detection by applying low-rank approximations to the weight matrices of CLIP.

6. Conclusion

We introduced NegRefine, a novel framework for zero-shot OOD detection, addressing key limitations of previous negative-label-based methods. NegRefine integrates a negative label filtering mechanism and a multi-matching-aware scoring function. The filtering mechanism refines the negative label set by removing subcategory labels and proper nouns, ensuring that the negative set has lower similarity to in-distribution samples. The scoring function dynamically adjusts the contribution of in-distribution labels, mitigating the incorrect detection of in-distribution samples as OOD caused by the multi-label matching issue. Extensive evaluations on ImageNet-1K benchmark with reliable OOD data demonstrate that NegRefine outperforms state-of-the-art methods on most OOD datasets.

Acknowledgments

This project was supported in part by a discovery grant for Ke Wang from Natural Sciences and Engineering Research Council of Canada, and by collaborative research funding from the National Research Council of Canada's Artificial Intelligence for Logistics Program.

References

[1] Amirhossein Ansari, Ke Wang, and Pulei Xiong. Out-of-distribution aware classification for tabular data. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 65–75, 2024. 1

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 5

[3] Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *ICML*, 2023. 2, 4, 5

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 7

[5] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024. 3

[6] Mengyuan Chen, Junyu Gao, and Changsheng Xu. Conjugated semantic pool improves OOD detection with pre-trained vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2, 3, 4, 5, 7, 8

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 4

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 4

[9] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *Proceedings of the International Conference on Learning Representations*, 2022. 1, 5

[10] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6568–6576, 2022. 1, 2, 5, 7

[11] Christiane Fellbaum. *WordNet: An electronic lexical database*. MIT press, 1998. 1, 3, 8

[12] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017. 1

[13] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 7

[14] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021. 7

[15] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021. 4

[16] Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided ood detection with pretrained vision-language models. In *International Conference on Learning Representations (ICLR)*, 2024. 1, 2, 3, 4, 5, 7, 8

[17] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 7

[18] Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning transferable negative prompts for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17584–17594, 2024. 8

[19] Yixia Li, Boya Xiong, Guanhua Chen, and Yun Chen. SeTAR: Out-of-distribution detection with selective low-rank approximation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 8

[20] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 1

[21] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020. 1

[22] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35:35087–35102, 2022. 1, 2, 5, 7

[23] Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10051–10059, 2022. 7

[24] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. *Advances in Neural Information Processing Systems*, 36:76298–76310, 2023. 8

[25] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Gl-mcm: Global and local maximum concept matching for zero-shot out-of-distribution detection. *International Journal of Computer Vision*, pages 1–11, 2025. 2, 5, 6, 7

[26] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023. 3

[27] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015. 1

[28] Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. Out-of-distribution detection with negative prompts. In *The Twelfth International Conference on Learning Representations*, 2024. 8

[29] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 7

[30] 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*, pages 8748–8763. PMLR, 2021. 1, 3, 5

[31] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 7

[32] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 7

[33] Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. 8

[34] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 4, 5

[35] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 7

[36] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 7

[37] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4921–4930, 2022. 5

[38] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812, 2023. 1, 2, 4, 5, 7

[39] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, pages 23631–23644. PMLR, 2022. 1

[40] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 4

[41] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024. 1

[42] Yabin Zhang and Lei Zhang. Adaneg: Adaptive negative proxy guided OOD detection with vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 8

[43] Yabin Zhang, Wenjie Zhu, Chenhang He, and Lei Zhang. Lapt: Label-driven automated prompt tuning for ood detection with vision-language models. In *European Conference on Computer Vision*, pages 271–288. Springer, 2024. 8

[44] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 4

[45] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 6