S-CLIP: Semi-supervised Vision-Language Pre-training using Few Specialist Captions

Sangwoo Mo **Minkyu Kim** **Kyungmin Lee** **Jinwoo Shin**
KAIST
{swmo, minkyu.kim, kyungmnlee, jinwoos}@kaist.ac.kr

Abstract

Vision-language models, such as contrastive language-image pre-training (CLIP), have demonstrated impressive results in natural image domains. However, these models often struggle when applied to specialized domains like remote sensing, and adapting to such domains is challenging due to the limited number of image-text pairs available for training. To address this, we propose S-CLIP, a semi-supervised learning method for training CLIP that utilizes additional unpaired images. S-CLIP employs two pseudo-labeling strategies specifically designed for contrastive learning and the language modality. The caption-level pseudo-label is given by a combination of captions of paired images, obtained by solving an optimal transport problem between unpaired and paired images. The keyword-level pseudo-label is given by a keyword in the caption of the nearest paired image, trained through partial label learning that assumes a candidate set of labels for supervision instead of the exact one. By combining these objectives, S-CLIP significantly enhances the training of CLIP using only a few image-text pairs, as demonstrated in various specialist domains, including remote sensing, fashion, scientific figures, and comics. For instance, S-CLIP improves CLIP by 10% for zero-shot classification and 4% for image-text retrieval on the remote sensing benchmark, matching the performance of supervised CLIP while using three times fewer image-text pairs.[1]

1 Introduction

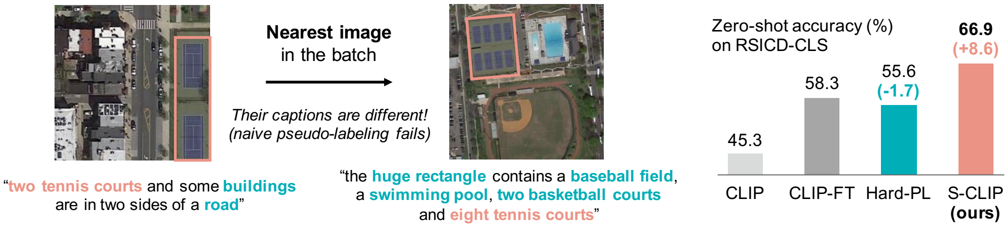
Pre-trained vision-language models have achieved remarkable success, providing a foundation for numerous downstream tasks [1]. However, these models often struggle when applied to specialized domains such as remote sensing or medical imaging [2]. This is because the models are trained on web-crawled data that may not fully capture the diversity and complexity of these domains [3, 4]. To address this issue, previous research has focused on constructing large-scale pre-training datasets for each domain [5, 6]. However, annotating caption for each specialized domain can be expensive and time-consuming, which limits their applicability across various domains.

Several studies have attempted to reduce the number of image-text pairs used for vision-language pre-training. However, these approaches often rely on additional information, such as pre-trained object detectors [7–9], or class-annotated images [10], which may not be applicable to specialized domains. Other approaches leverage self-supervised learning to utilize unpaired data [11–14], but they do not fully exploit the information provided by image-text pairs. Research on leveraging a few image-text pairs to improve vision-language pre-training remains underexplored.

In various areas of machine learning, particularly in image classification, semi-supervised learning [15] is a popular approach for training a model using limited annotations. This technique typically relies on pseudo-labeling [16], which predicts labels for unlabeled data by propagating information

[1] Code: https://github.com/alinlab/s-clip

Preprint. Under review.



(a) Problem of naive pseudo-labeling (use the nearest caption)

(b) Comparison of models

Figure 1: **Motivation.** (a) Images and captions are from the remote sensing dataset, RSICD. Unlike class labels, captions vary widely across images, making the naive pseudo-labeling of assigning the nearest caption ineffective. (b) Zero-shot accuracy on the RSICD-CLS dataset. Bracket indicates the gap from CLIP fine-tuned on remote sensing datasets, called CLIP-FT. The original CLIP struggles in this specialized domain, and CLIP-FT gives a large gain. As discussed in (a), naive pseudo-labeling (Hard-PL) harms CLIP-FT, while our S-CLIP provides substantial improvement.

from labeled data and refining the model using them. Combining this approach with self-supervised learning [17] has significantly improved classifiers using only a few class labels [18–22].

A natural question is whether semi-supervised learning, particularly pseudo-labeling, can improve vision-language pre-training. However, the techniques from image classification cannot be directly applied to vision-language models since captions are diverse and often uniquely associated with each image. Thus, the naive pseudo-labeling approach of assigning the caption of the nearest labeled image[2] (Hard-PL) [16] can mislead the models. Figure 1a demonstrates that even visually similar images often have different captions, although they may share keywords such as "tennis court." As a result, Figure 1b shows that Hard-PL harms the performance of CLIP fine-tuning. This observation motivates the proper design of pseudo-labeling method that effectively utilizes captions.

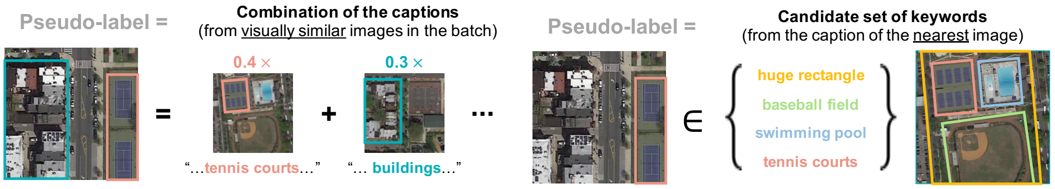
Contribution. We propose S-CLIP, a novel semi-supervised learning method for vision-language pre-training, particularly CLIP [23]. S-CLIP introduces two novel pseudo-labeling methods specifically designed for contrastive learning and the language modality, as illustrated in Figure 2:

- *Caption-level pseudo-label.* We assume that the semantics of an unlabeled image can be expressed as a combination of those of labeled images. To achieve this, we define the caption-level pseudo-label as a probability distribution over the labeled images, which guides the relationship of an unlabeled image with its corresponding captions. Specifically, the pseudo-labels are obtained by solving an optimal transport [24, 25] problem between the unlabeled and labeled images. This approach prevents pseudo-labels from collapsing to a few captions and ensures robust training, particularly when dealing with distribution shifts between unlabeled and labeled images.
- *Keyword-level pseudo-label.* We assume that an unlabeled image shares keywords with visually similar images, even if their full captions are not identical. Therefore, we define the keyword-level pseudo-label as one of the keywords in the nearest labeled image to an unlabeled image. This approach creates a candidate set of target keywords instead of a single exact one, and the training can be formulated as a partial label learning [26] problem. This loss helps the model understand the local components of unlabeled images, leveraging the structure of language.

We note that both pseudo-labels are complementary. Caption-level pseudo-label helps the model understand the global structure of language, which is more effective for image-text retrieval. Keyword-level pseudo-label helps the model understand the local phrases, which is more effective for zero-shot classification. Combining them achieves the best of both worlds (Section 5.4).

We demonstrate the effectiveness of our method in various specialist domains with limited available image-text pairs, including remote sensing [27], fashion [28], scientific figures [29], and comics [30] domains. In the remote sensing domain, S-CLIP outperforms CLIP fine-tuning and semi-supervised learning competitors in five zero-shot classification and six image-text retrieval tasks. For instance, S-CLIP improves zero-shot accuracy on WHU-RS19 [31] by 10.4% and image-to-text retrieval R@5 on UCM [32] by 4.4% compared to CLIP fine-tuning. S-CLIP remains robust even when unlabeled images are drawn from a different dataset. In the comics domain, the Simpsons [30] dataset only

[2] We use the terms "paired" and "labeled" interchangeably, as well as "unpaired" and "unlabeled," depending on the context of vision-language pre-training or semi-supervised learning.



(a) Caption-level pseudo-label

(b) Keyword-level pseudo-label

Figure 2: **Method overview.** Conceptual illustration of our proposed S-CLIP, utilizing (a) caption-level and (b) keyword-level pseudo-labels. In (a), the pseudo-label is given by a probability distribution over the captions, obtained by solving an optimal transport problem between unlabeled and labeled images. In (b), the pseudo-label is given by the candidate set of keywords from the caption of the nearest labeled image, which is learned by the partial label learning algorithm.

includes 800 image-text pairs. Therefore, we incorporate the Simpsons Characters [33] dataset, which consists of 30,000 images without captions. Here, S-CLIP improves the text→image retrieval R@1 of CLIP from 11.8% to 15.8%, providing a relative gain of 33%.

2 Related Work

Vision-language pre-training (VLP). Pre-training vision-language models has achieved remarkable success, providing a foundation for many downstream tasks [1]. Numerous approaches have been proposed, including reconstructing masked inputs [34–38], learning a joint embedding of vision and language modalities [23, 39–42], generating language descriptions from image inputs [43–45], and connecting pre-trained unimodal image and language models [46–48]. CLIP [23] is a representative model that learns a joint embedding between vision and language modalities. CLIP has demonstrated effectiveness in various applications, including zero-shot classification [23], image-text retrieval [49], open-vocabulary object segmentation [50, 51], and out-of-distribution detection [52].

VLP for specialist domains. Despite their success in natural image domains, VLP models often struggle when applied to specialist domains like remote sensing [2]. This is because they are trained on web-crawled data that may not capture the diversity and complexity of these domains [3, 4]. Data in these fields is often scarce, has limited accessibility, and requires expert annotations, making it difficult to obtain the image-text pairs needed for VLP training. While attempts have been made to train VLP models for specialized domains, including medical [10, 53–59], fashion [60–63], and remote sensing [64], previous research has mainly focused on constructing large-scale pre-training datasets for each domain [5, 6], which can be expensive and time-consuming.

VLP with limited pairs. Several studies trained VLP models with a limited number of image-text pairs. Hsu et al. [65] used adversarial domain adaptation to regularize unpaired data embeddings similar to paired data. However, this approach can cause problems when the unpaired data is drawn from a different distribution. MedCLIP [10] utilized images with class labels to reduce the need for captions. Nevertheless, it still requires class labels, limiting the use of large unlabeled data. Other studies explored unsupervised or semi-supervised VLP training in image domains [7–9]. However, they rely on a pre-trained object detector to align detected objects with keywords in captions, making them unsuitable for specialist domains. Other works leverage self-supervised learning on each modality [11–14], but do not fully exploit the information provided by image-text pairs. In contrast, our method leverages image-text pairs and is generally applicable without extra information.

Semi-supervised learning. Semi-supervised learning [15] aims to train a model using a small labeled dataset and a large unlabeled dataset. Techniques for semi-supervised learning can be categorized into two types: pseudo-labeling (or label propagation) [16, 66], which utilizes information from the small labeled dataset, and self-supervised learning [17], which learns representations from the unlabeled dataset. Combining these techniques has achieved remarkable success in semi-supervised learning, particularly for training classifiers with few class labels [18–22]. Prior works on robust semi-supervised learning focus on scenarios where the distributions of unlabeled and labeled data shift [67–69]. This issue naturally arises in vision-language pre-training due to the uniqueness of captions for each image. Our method addresses this issue using soft labels and performs robustly even when the unlabeled images are from a different dataset than the labeled images.

3 Background

Before we demonstrate our semi-supervised vision-language pre-training method, we provide a brief overview on contrastive language-image pre-training and semi-supervised learning.

3.1 Contrastive language-image pre-training

Contrastive language-image pre-training (CLIP) [23] is a popular vision-language model that learns a joint embedding space connecting images x and texts y.[3] CLIP is trained using a batch of paired images and texts $\{x_i, y_i\}_{i=1}^N$ through contrastive learning, where those from the same pair are treated as positive and different pairs as negative. The training objective of CLIP involves two classification tasks: predicting a text given an image $p(y|x)$ and predicting an image given a text $p(x|y)$. Labels are assigned to the N samples in the batch, corresponding to their pairs as the target. The embeddings are normalized to have a unit norm, and CLIP predicts the label using a softmax function with a temperature parameter $\tau > 0$, applied to the cosine similarity of the embeddings.

Specifically, we denote the softmax classifier σ_τ as a function of input and target embeddings. This allows us to represent $p(y|x) = \sigma_\tau(x, \{y_i\}_{i=1}^N) \in \mathbb{R}^N$ and $p(x|y) = \sigma_\tau(y, \{x_i\}_{i=1}^N) \in \mathbb{R}^N$, which denote the class probabilities of CLIP. These probabilities are calculated as follows: $p(y = y_i|x) = \exp(x \cdot y_i/\tau)/(\sum_{j=1}^N \exp(x \cdot y_j/\tau))$. The CLIP model minimizes the following objective:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2N} \sum_{i=1}^{N} \left(H(p(y|x_i), \mathbf{e}_i) + H(p(x|y_i), \mathbf{e}_i) \right), \tag{1}$$

where H denotes the cross-entropy loss and $\mathbf{e}_i \in \mathbb{R}^N$ is a one-hot vector with the i-th element being one. The CLIP model can be used for zero-shot classification or image-text retrieval by computing cosine similarity in the embedding space. For example, in zero-shot classification, the text embedding of the class prompt that is closest to an image embedding can be identified.

3.2 Semi-supervised learning

Semi-supervised learning (Semi-SL) [15] aims to train a model using a small number of labeled data $\{x_i, y_i\}_{i=1}^N$ and a large number of unlabeled data $\{u_i\}_{i=1}^M$, i.e., when $M \gg N$. Various Semi-SL methods have been proposed to enhance neural network performance, particularly for image classification [19–22].[4] These methods typically rely on pseudo-labels [16], obtained by predicting classes for unlabeled data using a model. The model is then refined by minimizing the cross-entropy loss $\frac{1}{M} \sum_{i=1}^M H(p(y|u_i), q_i)$ with respect to the prediction $p(y|u_i) \in \mathbb{R}^C$ and pseudo-label $q_i \in \mathbb{R}^C$ for unlabeled data u_i. Pseudo-labels can take the form of a hard label, represented by a one-hot vector identifying the class with the highest probability [16], or a soft label that calibrates this hard label [21]. Pseudo-labels minimize prediction entropy by sharpening specific classes [70].

Semi-SL for CLIP. Assigning pseudo-labels in CLIP poses challenges due to the unique captions associated with each image. The assumption that unlabeled data belongs to the same class as labeled data does not hold here. Therefore, we propose a novel Semi-SL method for CLIP.

4 S-CLIP: Semi-supervised Vision-Language Pre-training

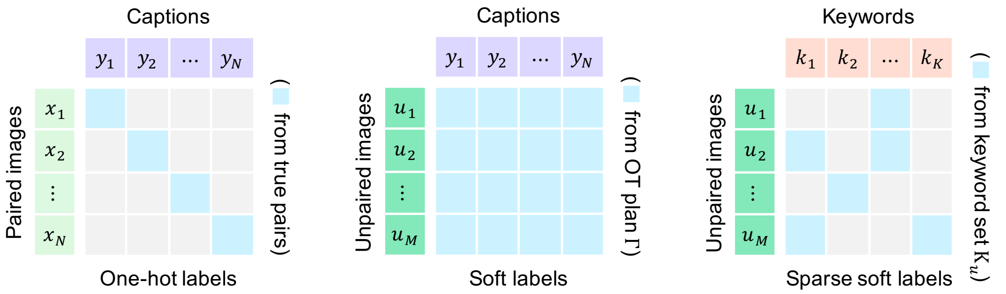
We extend the training of CLIP to incorporate unpaired images $\{u_i\}_{i=1}^M$ alongside the image-text pairs $\{x_i, y_i\}_{i=1}^N$. To achieve this, we introduce S-CLIP, which integrates two novel pseudo-labeling methods considering contrastive learning and the language modality. The conceptual illustrations of the proposed pseudo-labels and training objectives are shown in Figure 2 and Figure 3, respectively.

4.1 Caption-level pseudo-label

Although the semantics of an unlabeled image u may not precisely match any captions y_i, visually similar images often share similar semantics. Therefore, the caption for unlabeled image u can be expressed as a combination of the captions $\{y_i\}$. To achieve this, we use caption-level pseudo-labels,

[3] We use the notations x and y to represent image and text embeddings computed by the CLIP encoders. However, for simplicity, we often omit the term "embeddings" when referring to them.

[4] In image classification, the targets y are represented by one-hot vectors of size C for C classes.



(a) CLIP loss $\mathcal{L}_{\text{CLIP}}$ (b) Caption loss $\mathcal{L}_{\text{caption}}$ (c) Keyword loss $\mathcal{L}_{\text{keyword}}$

Figure 3: **Training objective.** Conceptual illustrations of the training objectives of S-CLIP: (a) is the original CLIP loss in Eq. (1) using paired images and texts; (b) and (c) are our proposed caption-level and keyword-level pseudo-label losses in Eq. (3) and Eq. (4), respectively, applied to unpaired images. Caption-level pseudo-labels are soft labels obtained from the optimal transport (OT) plan Γ, while keyword-level pseudo-labels are sparse soft labels obtained from the candidate keyword set K_u.

which represent a probability distribution over the captions. These pseudo-labels serve as the target for the N-way classification problem in contrastive learning, with a batch size of N. Figure 2a demonstrates how caption-level pseudo-label for an unlabeled image is obtained.

The pseudo-labels are derived from the relationship between unlabeled and labeled images, formulated as an optimal transport (OT) [24] problem. We define the cost function $\mathrm{C} \in \mathbb{R}^{M \times N}$, where M and N represent the number of unlabeled and labeled images, respectively. The cost is given by the negative cosine similarity between the normalized embeddings of the unlabeled image u_i and the labeled image x_j, i.e., $\mathrm{C}_{ij} = 1 - u_i \cdot x_j$. Then, we solve following entropic regularized OT [25] problem:

$$\min_{\Gamma \in \Pi(\mathbf{p},\mathbf{q})} \langle \Gamma, \mathrm{C} \rangle - \lambda H(\Gamma), \quad \Pi(\mathbf{p},\mathbf{q}) = \{\Gamma \in \mathbb{R}_+^{M \times N} \mid \Gamma \mathbf{1}_N = \mathbf{p},\ \Gamma^\top \mathbf{1}_M = \mathbf{q}\}. \tag{2}$$

Here, $\Pi(\mathbf{p},\mathbf{q})$ represents a set of transportation plans that satisfy the flow constraint. The sums of flows from sources and to sinks match the vectors $\mathbf{p} \in \mathbb{R}_+^M$ and $\mathbf{q} \in \mathbb{R}_+^N$, respectively, with both $\mathbf{p}$ and $\mathbf{q}$ adding up to one. We set them as uniform probabilities, and they worked in our experiments, even when unlabeled and labeled data have distribution shifts. $\mathbf{1}_N$ denotes the all-one vector of dimension N. The entropic regularizer $H(\Gamma) = -\sum_{i,j} \Gamma_{ij} \log \Gamma_{ij}$ with scale λ smooths the transportation plan, enabling an efficient solution using the Sinkhorn-Knopp [25] algorithm.

Once we obtain the transportation plan Γ, we define the pseudo-label for an unlabeled image u_i by normalizing its plan to sum up to one. Specifically, the j-th element of the pseudo-label $q_i \in \mathbb{R}^N$ is computed as $\Gamma_{ij} / \sum_j \Gamma_{ij}$. This pseudo-label is then utilized to guide the probability distribution over N captions, where the prediction $p(y|u_i) \in \mathbb{R}^N$ is determined by a softmax classifier $\sigma_\tau(u_i, \{y_i\}_{i=1}^N)$. The loss function for the caption-level pseudo-label aims to minimize the following:

$$\mathcal{L}_{\text{caption}} = \frac{1}{M} \sum_{i=1}^{M} H(p(y|u_i), q_i). \tag{3}$$

The prediction $p(y|u_i)$ relates the unlabeled image u_i to the text embeddings y, while the pseudo-label q_i relates it to the image embeddings x. Consequently, the proposed loss function aligns the relationships between image and text embeddings, as suggested in supervised CLIP [71–73]. In our experiments, we compare the usage of text and image embeddings for pseudo-label computation, and using image embeddings outperformed using text embeddings due to their robustness in novel specialist domains. We also explore computing pseudo-embeddings [9], but it yields inferior results as synthesizing pseudo-embeddings is more challenging than inferring pseudo-labels.

Effect of OT. Optimal transport balances the flow between unlabeled and labeled images, ensuring robust pseudo-label assignments without concentration on a few labels. In vision-language tasks, diverse and imbalanced captions often cause image embeddings to focus on a few nearest captions, resulting in the collapse of embeddings. This challenge is magnified when employing semi-supervised learning with distribution shifts [67]. We empirically validate the detrimental effect of naive pseudo-labeling approaches, especially in the presence of distribution shifts.

Compared to soft-nearest neighbor pseudo-labels (Soft-PL) [21], our method achieves balance through Sinkhorn iterations. Specifically, it reverts to Soft-PL when the iteration count is zero. In our experiments, we use 10 iterations, and increasing the iteration count has only a marginal impact on performance. The computational cost is negligible as the process occurs in a low-dimensional embedding space. We set the scale of the entropic regularizer to match the softmax temperature, $\lambda = \tau$. This adjustment scales the cost function by $1/\tau$, aligning it with the cosine similarity scale learned by CLIP. More information on the OT and Soft-PL relationship is in Appendix A.

OT for vision-language models. Several works have utilized OT for vision-language models, but for problems other than semi-supervised learning. We discuss these works in Appendix B.

4.2 Keyword-level pseudo-label

We propose a keyword-level pseudo-labeling approach to overcome the limitations of caption-level pseudo-labels in capturing the meanings of words in captions. By leveraging the compositionality of language, we assume that visually similar images may share keywords, even if the entire captions differ. This enables us to guide unlabeled images using keywords from their nearest neighbors. It is important to note that an image may have multiple keywords, and visually similar images share only a subset. This aligns with the partial label learning (PLL) [26] problem, where the target label is unknown, and we have access to a candidate set of labels. For instance, in Figure 1a, the image shares the keyword "tennis court" with its nearest neighbor but not "baseball field." Consequently, the nearest neighbor provides a candidate set of keywords, as illustrated in Figure 2b.

To identify candidate keywords for an unlabeled image u, we assume a pre-defined set of keywords with corresponding embeddings $\mathsf{K} = \{k_i\}_{i=1}^{K}$. These keywords can be obtained from class names, if available, or extracted by using algorithms like YAKE [74] applied to captions. Given an unlabeled image u, we find the nearest labeled image x using the OT assignments from the previous section. Let $\{c_1, \ldots, c_l\}$ be the indices of keywords present in labeled image x, denoted as $\mathsf{K}_u = \{k_{c_1}, \ldots, k_{c_l}\}$. This forms the candidate set of target keywords for u. While we simply use the keywords explicitly from the caption, some additional steps can be taken to infer synonyms or related concepts.

In the context of partial label learning, we assume that the ground-truth target label for the unlabeled image u belongs to the candidate set K_u. PLL aims to minimize the loss by selecting the minimum value among candidate labels, $\min_{k \in \mathsf{K}_u} H(p(k|u), k)$, rather than specifying a single target label. However, relying on a single minimum can be risky. Previous works have addressed this by using soft labels for candidates [75, 76]. This involves minimizing $H(p(k|u), q)$, where $q \in \mathbb{R}^K$ is a soft label with positive values for keywords in the candidate set and zero for others.

Similarly, we define the pseudo-label q_i for the unlabeled image u_i based on the similarity between the embeddings of u_i and the keywords k_j in the candidate set K_u. Specifically, the j-th element of the pseudo-label q_i is calculated using a softmax function $\sigma_\tau(u_i, \mathsf{K}_u)$ for $k \in \mathsf{K}_u$ and is set to zero otherwise. This pseudo-label is then utilized to guide the probability distribution over K keywords, where the prediction $p(k|u) \in \mathbb{R}^K$ is determined by a softmax classifier $\sigma_\tau(u_i, \mathsf{K})$. The loss function for the keyword-level pseudo-label aims to minimize the following:

$$\mathcal{L}_{\text{keyword}} = \frac{1}{M} \sum_{i=1}^{M} H(p(k|u_i), q_i). \tag{4}$$

Both the prediction $p(k|u_i)$ and the pseudo-label q_i are related to the unlabeled image and keywords. However, the restricted candidate set effectively guides the learning process through PLL.

Training objective. We apply the CLIP loss to paired images and texts and the proposed pseudo-label losses to unpaired images. The pseudo-label losses are halved to match the scale of the image part in the CLIP loss. In summary, our final training objective becomes $\mathcal{L}_{\text{CLIP}} + {}^{1}\!/_{2}\,(\mathcal{L}_{\text{caption}} + \mathcal{L}_{\text{keyword}})$.

5 Experiments

We demonstrate the effectiveness of our method in various specialist domains with limited available image-text pairs, including remote sensing, fashion, scientific figures, and comics domains.

Setup. We use a pre-trained CLIP model on natural images [23], called CLIP (original). However, this model often struggles in specialist domains. Therefore, we fine-tune the model with domain-specific image-text pairs, called CLIP (fine-tune). We compare this supervised model with semi-supervised

Table 1: Zero-shot classification results on remote sensing datasets. We compare the original CLIP, supervised CLIP fine-tuned on labeled data (L), and semi-supervised methods that utilize unlabeled data sampled from the same (L=U) or different (L≠U) distribution as the labeled data. Parentheses indicate the performance gap from the supervised CLIP, where values highlighted in green indicate gaps larger than one. Bolds denote the best results among the semi-supervised methods within the same setups. S-CLIP consistently improves zero-shot accuracy, even when the unlabeled data has a distribution shift. In contrast, naive pseudo-labeling often harms the accuracy.

Method	Data	RSICD-CLS	UCM-CLS	WHU-RS19	RSSCN7	AID
CLIP (original)	-	45.3	50.5	65.5	58.9	47.8
CLIP (fine-tune)	L	58.3±0.3	63.5±3.4	76.5±3.2	61.9±1.2	63.1±1.3
Hard-PL [16]	L=U	56.6±3.5 (-1.7)	61.6±2.2 (-1.9)	78.1±2.5 (+1.6)	63.9±2.1 (+2.0)	63.2±2.6 (+0.1)
Soft-PL [21]		62.5±0.8 (+4.2)	65.7±2.7 (+2.2)	83.7±2.7 (+7.2)	65.7±0.6 (+3.8)	68.0±0.7 (+4.9)
S-CLIP (ours)		**66.9**±1.7 (+8.6)	**66.7**±1.6 (+3.2)	**86.9**±2.0 (+10.4)	**66.2**±1.1 (+4.3)	**73.0**±0.3 (+9.9)
Hard-PL [16]	L≠U	55.8±1.2 (-2.5)	61.0±6.2 (-2.5)	76.3±1.1 (-0.2)	62.5±2.7 (+0.6)	62.3±1.4 (-0.8)
Soft-PL [21]		56.1±3.5 (-2.2)	62.4±1.8 (-1.1)	79.5±3.6 (+3.0)	63.3±2.6 (+1.4)	62.4±2.1 (-0.7)
S-CLIP (ours)		**65.3**±2.1 (+7.0)	**66.0**±0.4 (+2.5)	**86.3**±1.1 (+9.8)	**68.3**±2.2 (+6.4)	**70.8**±2.5 (+7.7)

Table 2: Image-text retrieval results on remote sensing datasets, following the same setup of Table 1. Bolds denote the best results among the semi-supervised methods within the same setups. S-CLIP performs best in most cases, except the supervised baseline works best for text→image retrieval on the Sydney dataset. Naive pseudo-labeling also helps, but the gain is often unstable.

		Image→text R@5			Text→image R@5		
Method	Data	RSICD	UCM	Sydney	RSICD	UCM	Sydney
CLIP (original)	-	9.4	34.3	36.2	10.1	24.8	51.7
CLIP (fine-tune)	L	15.4±1.7	41.3±1.8	47.1±6.5	15.1±1.0	40.9±1.6	56.1±2.4
Hard-PL [16]	L=U	16.1±0.2	40.8±2.9	43.1±3.0	15.7±0.7	40.5±3.0	47.7±5.3
Soft-PL [21]		17.0±0.9	43.2±3.9	42.0±4.3	16.5±0.1	42.9±3.3	50.2±4.9
S-CLIP (ours)		**18.4**±0.6	**45.7**±1.4	**50.0**±3.0	**16.8**±1.2	**43.5**±1.5	**55.1**±2.0
Hard-PL [16]	L≠U	15.8±1.0	42.1±5.7	47.1±4.0	15.3±0.4	42.1±3.1	50.0±1.7
Soft-PL [21]		16.5±1.1	40.2±3.2	46.6±4.6	15.4±0.2	40.8±0.5	**54.0**±4.0
S-CLIP (ours)		**17.1**±0.8	**43.5**±3.5	**48.9**±2.6	**15.8**±0.7	**42.5**±1.5	52.3±1.0

methods using unpaired images. We use a batch size of 64 per GPU, with a total of 4 GPUs. To ensure fair GPU memory usage in semi-supervised learning, we employ 32 image-caption pairs and 32 unpaired images. Models are evaluated on zero-shot classification and image-text retrieval tasks, measuring Top-1 classification accuracy (%) and recall at K (R@K). We report the average and standard deviation across three random seeds. Additional details are in Appendix D.

Baselines. We compare S-CLIP with the semi-supervised extension of CLIP using pseudo-labels: hard one-hot (Hard-PL) [16] or soft nearest neighbor (Soft-PL) [21]. See Appendix A for details.

5.1 Remote sensing datasets

We train vision-language models using the union of RSICD [27], UCM [32], and Sydney [77], named RS-ALL, following the setup of [64]. For semi-supervised learning, we subsample 10% of image-text pairs as labeled data (L), while the remaining 90% of images (L=U) or unlabeled images from the RESISC45 [78] dataset (L≠U) served as unlabeled data. In Appendix E, we demonstrate that our method consistently gives improvements over CLIP across various setups. These include different neural architectures, varied ratios of image-text pairs, and even scenarios where we use all image-text pairs from RS-ALL in conjunction with unlabeled images from RESISC45.

For zero-shot classification, we use the validation sets of the classification versions of the RSICD and UCM datasets, denoted as RSCID-CLS and UCM-CLS, respectively. To evaluate the generalization abilities, we test models on unseen datasets such as WHU-RS19 [31], RSSCN7 [79], and AID [80]. For image-text retrieval, we use the validation sets of the RSICD, UCM, and Sydney datasets.

Zero-shot classification. Table 1 presents the zero-shot classification results. S-CLIP consistently and significantly outperforms all supervised CLIP fine-tuning and semi-supervised methods. One can make several observations. Firstly, fine-tuning CLIP improves performance over the original

Table 3: Zero-shot classification results on fashion datasets. Parentheses indicate the performance gap from the supervised CLIP, where values highlighted in green indicate gaps larger than one. Bolds denote the best results. S-CLIP consistently outperforms all supervised and semi-supervised methods.

	Fashion200k		FashionGen		Polyvore
Method	Super-class	Sub-class	Super-class	Sub-class	Class
CLIP (original)	73.4	29.3	35.9	22.1	58.4
CLIP (fine-tune)	79.4±1.5	38.4±1.5	41.5±1.4	33.1±1.9	64.2±3.1
Hard-PL [16]	74.6±2.7 (-4.8)	35.0±2.6 (-3.4)	36.6±3.9 (-4.9)	28.8±0.8 (-4.3)	65.5±0.7 (+1.3)
Soft-PL [21]	80.0±1.1 (+0.6)	37.1±0.7 (-1.3)	45.5±3.2 (+4.0)	33.2±0.3 (+0.1)	73.6±7.9 (+9.4)
S-CLIP (ours)	**82.0**±2.8 (+2.6)	**39.5**±1.1 (+1.1)	**55.3**±2.6 (+13.8)	**38.7**±1.9 (+5.6)	**74.6**±1.8 (+10.4)

Table 4: Image-text retrieval results on SciCap (scientific figures) and Simpsons (comics) datasets. We train the model on each dataset and evaluate on their validation sets. Bolds denote the best results. S-CLIP consistently outperforms all supervised and semi-supervised methods.

	SciCap				Simpsons			
	Image→text		Text→image		Image→text		Text→image	
Method	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
CLIP (original)	9.2	14.7	9.0	14.5	15.8	40.8	10.5	27.6
CLIP (fine-tune)	10.5±0.2	17.9±0.2	10.2±0.1	17.2±0.2	16.7±2.0	41.4±1.1	11.8±2.3	41.7±1.5
Hard-PL [16]	9.7±0.3	16.8±0.6	9.2±0.6	16.2±0.7	15.4±1.5	41.7±3.3	12.7±1.5	40.8±2.3
Soft-PL [21]	9.1±0.5	15.5±1.0	9.7±0.3	16.8±0.5	15.4±2.0	39.9±2.7	12.9±1.9	36.4±2.7
S-CLIP (ours)	**10.9**±0.3	**18.8**±0.3	**11.1**±0.3	**18.8**±0.4	**18.4**±2.3	**43.9**±1.5	**15.8**±1.3	**42.1**±0.1

CLIP, highlighting the importance of adapting models to the target specialist domain. Secondly, our proposed pseudo-labeling techniques are crucial, as naive pseudo-labeling often harms accuracy. Thirdly, semi-supervised learning methods improve performance on unseen datasets, such as RSSCN7, demonstrating the usefulness of observing more unlabeled images in generalization. Lastly, our method is robust to the distribution shifts (L≠U) setup, which leverages an external source of unlabeled images, a common scenario in practice. In contrast, Soft-PL improves performance in the same distribution (L=U) setup but fails on the distribution shifts scenario, confirming the effectiveness of optimal transport in balancing the assignments of pseudo-labels.

Image-text retrieval. Table 2 presents the image-text retrieval results. S-CLIP consistently improves both image→text and text→image retrieval, except in one case where the supervised baseline performs the best. This confirms that S-CLIP also benefits fine language understanding for retrieval. The trend is consistent across evaluation metrics, as seen in the R@1 results in Appendix E.4.

5.2 Fashion datasets

We train vision-language models using the union of Fashion200k [28], FashionGen [81], and Polyvore Outfits [82] datasets. We subsample 10% of the image-text pairs as labeled data and the remaining 90% of images as unlabeled data. We evaluate the zero-shot accuracy on the validation sets of all three datasets, considering both super-class and sub-class accuracies for Fashion200k and FashionGen. The class names in Polyvore match the same level of super-class as those in the other datasets.

Table 3 presents the zero-shot classification results. S-CLIP consistently outperforms all supervised and semi-supervised methods. The improvement is more significant for super-class classification but less pronounced for sub-class classification and image-text retrieval. This is because the current pseudo-labeling approach assumes that the semantics of unlabeled image can be represented by the captions in a batch. However, this assumption may not hold true for fine-grained semantics, as certain fine-grained captions may be missing from the batch. Therefore, our method could be further enhanced by increasing the batch size or incorporating a queue of caption embeddings [83, 84].

5.3 More captioning datasets

We conduct experiments on two more specialist domains, namely science figures and comics. For this purpose, we use the SciCap [29] and Simpsons [30] datasets. We train the models on each dataset

Table 5: Ablation study on design choices. We report the average values of zero-shot accuracy (left) and image↔text retrieval (right). Bolds denote the best results. See Section 5.4 for discussions.

(a) Training objective

Baseline	64.7	9.3
Caption-level	69.6	10.5
Keyword-level	70.0	9.9
Both	**71.9**	**10.6**

(b) Sinkhorn iteration

Iter. 0	70.6	9.2
Iter. 1	71.4	9.4
Iter. 5	71.6	10.2
Iter. 10	**71.9**	**10.6**

(c) Choice of keywords

YAKE-100	70.0	9.8
YAKE-200	**72.2**	9.3
YAKE-300	70.6	10.2
Class name	71.9	**10.6**

(d) Usage of OT plan

Pseudo-embed	68.8	9.2
Pseudo-label	**71.9**	**10.6**

(e) Embedding for OT

Text	71.6	9.7
Image	**71.9**	**10.6**

(f) Partial label loss

Hard-max	66.1	10.2
Soft-max	**71.9**	**10.6**

and evaluate their image-text retrieval performance on the validation sets. The results are presented in Table 4, and detailed discussions can be found in the following paragraphs.

SciCap. We subsample 10% of the image-text pairs as labeled data and the remaining 90% of images as unlabeled data. S-CLIP improves upon CLIP, even for the scientific figures domain, which exhibits a substantial gap from the natural image domain that CLIP is pre-trained on. This confirms the practical impact of S-CLIP in extending the applicability of CLIP to various specialist domains.

Simpsons. The Simpsons dataset contains only 800 image-text pairs, so we incorporate unlabeled images from another dataset called Simpsons Characters [33]. S-CLIP significantly improves CLIP, even in the presence of distribution shifts between unlabeled and labeled images. For example, it boosts the text→image retrieval R@1 from 11.8% to 15.8% (+33%). This confirms the practical impact of S-CLIP in specialist domains with limited available image-text pairs.

5.4 Ablation studies

We conduct an ablation study on the design choices for S-CLIP. The results are reported in Table 5, which compares the average values of zero-shot accuracy (across five tasks in Table 1) and image-text retrieval R@1 (across six tasks in Table 2) for each design. The observations are as follows:

(a) *Training objective.* Caption-level pseudo-labels are more beneficial for image-text retrieval, while keyword-level pseudo-labels are more beneficial for zero-shot classification. Combining both yields the best results, particularly in significantly improving zero-shot classification.

(b) *Sinkhorn iteration.* Sinkhorn iterations significantly improve the performance of Soft-PL (Iter. 0) by balancing pseudo-labels. However, it reaches saturation after 10 iterations. Thus, we use 10 iterations for our experiments, resulting in a negligible increase in computation costs.

(c) *Choice of keywords.* Both class names and keywords extracted by the YAKE [74] algorithm from the captions work well, although class names may provide more consistent improvement. This is further confirmed on the SciCap and Simpsons datasets, where class names are unavailable.

(d) *Usage of OT plan.* Using the transportation plan Γ for computing pseudo-embeddings, i.e., using $\sum_j q_{ij} \, y_j$ as the text embedding of an unlabeled image u_i, performs worse than pseudo-labeling. This is because synthesizing pseudo-embeddings requires more precise computation.

(e) *Embedding for OT.* Using image embeddings for computing pseudo-labels consistently outperforms using text embeddings, especially with distribution shifts between unlabeled and labeled images. This is because image embeddings are more robust in handling novel domains.

(f) *Partial label loss.* Using the proposed soft targets for partial label learning outperforms selecting the hard minimum from the candidate set, confirming the suggestion from prior works.

Further details on each point are discussed in Appendix F, along with additional ablation studies, including the trends in training curves and regularized fine-tuning approaches.

6 Conclusion

We propose S-CLIP, a semi-supervised extension of CLIP that leverages unpaired images to train in specialist domains with limited image-text pairs. We demonstrate its superiority across various domains such as remote sensing, fashion, scientific figures, and comics. We hope that S-CLIP expands the applicability of CLIP. Limitations and broader impacts are discussed in Appendix C.

References

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[2] Y. Kim, S. Mo, M. Kim, K. Lee, J. Lee, and J. Shin. Explaining visual biases as words by generating captions. *arXiv preprint arXiv:2301.11104*, 2023.

[3] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Conference of the Association for Computational Linguistics*, 2018.

[4] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[5] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Nature scientific data*, 2019.

[6] X. Yang, H. Zhang, D. Jin, Y. Liu, C.-H. Wu, J. Tan, D. Xie, J. Wang, and X. Wang. Fashion captioning: Towards generating accurate descriptions with semantic rewards. In *European Conference on Computer Vision*, 2020.

[7] L. H. Li, H. You, Z. Wang, A. Zareian, S.-F. Chang, and K.-W. Chang. Unsupervised vision-and-language pre-training without parallel images and captions. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2021.

[8] M. Zhou, L. Yu, A. Singh, M. Wang, Z. Yu, and N. Zhang. Unsupervised vision-and-language pre-training via retrieval-based multi-granular alignment. In *Conference on Computer Vision and Pattern Recognition*, 2022.

[9] T.-J. J. Wang, J. Laaksonen, T. Langer, H. Arponen, and T. E. Bishop. Learning by hallucinating: Vision-language pre-training with weak supervision. In *Winter Conference on Applications of Computer Vision*, 2023.

[10] Z. Wang, Z. Wu, D. Agarwal, and J. Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Conference on Empirical Methods in Natural Language Processing*, 2022.

[11] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2022.

[12] N. Mu, A. Kirillov, D. Wagner, and S. Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, 2022.

[13] E. Segal, B. Bogin, and J. Berant. Training vision-language models with less bimodal supervision. In *Automated Knowledge Base Construction*, 2022.

[14] R. Windsor, A. Jamaludin, T. Kadir, and A. Zisserman. Vision-language modelling for radiological imaging and reports in the low data regime. *arXiv preprint arXiv:2303.17644*, 2023.

[15] O. Chapelle, B. Scholkopf, and A. Zien. Semi-supervised learning. 2006. *Cambridge, Massachusettes: The MIT Press View Article*, 2, 2006.

[16] D.-H. Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 2013.

[17] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[18] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Neural Information Processing Systems*, 2019.

[19] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Neural Information Processing Systems*, 2020.

[20] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves imagenet classification. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[21] M. Assran, M. Caron, I. Misra, P. Bojanowski, A. Joulin, N. Ballas, and M. Rabbat. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *International Conference on Computer Vision*, 2021.

[22] M. Bošnjak, P. H. Richemond, N. Tomasev, F. Strub, J. C. Walker, F. Hill, L. H. Buesing, R. Pascanu, C. Blundell, and J. Mitrovic. Semppl: Predicting pseudo-labels for better contrastive representations. In *International Conference on Learning Representations*, 2023.

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[24] C. Villani et al. *Optimal transport: old and new*. Springer, 2009.

[25] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Neural Information Processing Systems*, 2013.

[26] T. Cour, B. Sapp, and B. Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 2011.

[27] X. Lu, B. Wang, X. Zheng, and X. Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 2017.

[28] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, pages 1463–1471, 2017.

[29] T.-Y. Hsu, C. L. Giles, and T.-H. Huang. Scicap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP*, 2021.

[30] D. Adler. Simpsons blip captions. https://huggingface.co/datasets/Norod78/simpsons-blip-captions, 2023.

[31] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maître. Structural high-resolution satellite image indexing. 2009.

[32] Y. Yang and S. Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *SIGSPATIAL international conference on advances in geographic information systems*, 2010.

[33] A. Attia. The simpsons characters data. https://www.kaggle.com/datasets/alexattia/the-simpsons-characters-dataset, 2018.

[34] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. Visualbert: A simple and performant baseline for vision and language. In *Annual Conference of the Association for Computational Linguistics*, 2020.

[35] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, 2020.

[36] W. Kim, B. Son, and I. Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 2021.

[37] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[38] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[39] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021.

[40] H. Pham, Z. Dai, G. Ghiasi, H. Liu, A. W. Yu, M.-T. Luong, M. Tan, and Q. V. Le. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*, 2021.

[41] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Conference on Computer Vision and Pattern Recognition*, 2022.

[42] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He. Scaling language-image pre-training via masking. *arXiv preprint arXiv:2212.00794*, 2022.

[43] K. Desai and J. Johnson. Virtex: Learning visual representations from textual annotations. In *Conference on Computer Vision and Pattern Recognition*, 2021.

[44] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

[45] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

[46] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill. Multimodal few-shot learning with frozen language models. In *Neural Information Processing Systems*, 2021.

[47] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Neural Information Processing Systems*, 2022.

[48] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[49] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.

[50] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022.

[51] M. Xu, Z. Zhang, F. Wei, Y. Lin, Y. Cao, H. Hu, and X. Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, 2022.

[52] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2023.

[53] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Conference on Computer Vision and Pattern Recognition*, 2018.

[54] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, 2022.

[55] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *International Conference on Computer Vision*, 2021.

[56] R. Liao, D. Moyer, M. Cha, K. Quigley, S. Berkowitz, S. Horng, P. Golland, and W. M. Wells. Multimodal representation learning via maximization of local mutual information. In *Medical Image Computing and Computer Assisted Intervention*, 2021.

[57] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European Conference on Computer Vision*, 2022.

[58] P. Müller, G. Kaissis, C. Zou, and D. Rueckert. Joint learning of localized representations from medical images and reports. In *European Conference on Computer Vision*, 2022.

[59] S. Zhang, Y. Xu, N. Usuyama, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023.

[60] D. Gao, L. Jin, B. Chen, M. Qiu, P. Li, Y. Wei, Y. Hu, and H. Wang. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

[61] M. Zhuge, D. Gao, D.-P. Fan, L. Jin, B. Chen, H. Zhou, M. Qiu, and L. Shao. Kaleido-bert: Vision-language pre-training on fashion domain. In *Conference on Computer Vision and Pattern Recognition*, 2021.

[62] S. Goenka, Z. Zheng, A. Jaiswal, R. Chada, Y. Wu, V. Hedau, and P. Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Conference on Computer Vision and Pattern Recognition*, 2022.

[63] X. Han, L. Yu, X. Zhu, L. Zhang, Y.-Z. Song, and T. Xiang. Fashionvil: Fashion-focused vision-and-language representation learning. In *European Conference on Computer Vision*, 2022.

[64] A. Arampacha, V. Dev, G. Venkatesh, B. Mayank, G. Ritobrata, and P. Sujit. Fine tuning clip with remote sensing (satellite) images and captions. https://huggingface.co/blog/fine-tune-clip-rsicd, 2021.

[65] T.-M. H. Hsu, W.-H. Weng, W. Boag, M. McDermott, and P. Szolovits. Unsupervised multimodal representation learning across medical images and reports. In *NeurIPS Machine Learning for Health Workshop*, 2018.

[66] Y. Bengio, O. Delalleau, and N. Le Roux. label propagation and quadratic criterion. 2006.

[67] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Neural Information Processing Systems*, 2018.

[68] J. Park, S. Yun, J. Jeong, and J. Shin. Opencos: Contrastive semi-supervised learning for handling open-set unlabeled data. *arXiv preprint arXiv:2107.08943*, 2021.

[69] S. Mo, J.-C. Su, C.-Y. Ma, M. Assran, I. Misra, L. Yu, and S. Bell. Ropaws: Robust semi-supervised representation learning from uncurated data. In *International Conference on Learning Representations*, 2023.

[70] Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In *Neural Information Processing Systems*, 2004.

[71] J. Zhou, L. Dong, Z. Gan, L. Wang, and F. Wei. Non-contrastive learning meets language-image pre-training. *arXiv preprint arXiv:2210.09304*, 2022.

[72] S. Goel, H. Bansal, S. Bhatia, R. Rossi, V. Vinay, and A. Grover. Cyclip: Cyclic contrastive language-image pretraining. In *Neural Information Processing Systems*, 2022.

[73] Y. Gao, J. Liu, Z. Xu, T. Wu, W. Liu, J. Yang, K. Li, and X. Sun. Softclip: Softer cross-modal alignment makes clip stronger. *arXiv preprint arXiv:2303.17561*, 2023.

[74] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509: 257–289, 2020.

[75] J. Lv, M. Xu, L. Feng, G. Niu, X. Geng, and M. Sugiyama. Progressive identification of true labels for partial-label learning. In *International Conference on Machine Learning*, 2020.

[76] H. Wang, R. Xiao, Y. Li, L. Feng, G. Niu, G. Chen, and J. Zhao. Pico: Contrastive label disambiguation for partial label learning. In *International Conference on Learning Representations*, 2022.

[77] F. Zhang, B. Du, and L. Zhang. Saliency-guided unsupervised feature learning for scene classification. *IEEE transactions on Geoscience and Remote Sensing*, 2014.

[78] G. Cheng, J. Han, and X. Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 2017.

[79] Q. Zou, L. Ni, T. Zhang, and Q. Wang. Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 2015.

[80] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2017.

[81] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, and C. Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018.

[82] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis. Learning fashion compatibility with bidirectional lstms. In *ACM international conference on Multimedia*, 2017.

[83] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020.

[84] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[85] S. Yuan, K. Bai, L. Chen, Y. Zhang, C. Tao, C. Li, G. Wang, R. Henao, and L. Carin. Weakly supervised cross-domain alignment with optimal transport. In *British Machine Vision Conference*, 2020.

[86] G. Chen, W. Yao, X. Song, X. Li, Y. Rao, and K. Zhang. Prompt learning with optimal transport for vision-language models. In *International Conference on Learning Representations*, 2023.

[87] K. Kim, Y. Oh, and J. C. Ye. Zegot: Zero-shot segmentation through optimal transport of text prompts. *arXiv preprint arXiv:2301.12171*, 2023.

[88] T. Gupta, A. Vahdat, G. Chechik, X. Yang, J. Kautz, and D. Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, 2020.

[89] B. Wu, R. Cheng, P. Zhang, P. Vajda, and J. E. Gonzalez. Data efficient language-supervised zero-shot recognition with optimal transport distillation. In *International Conference on Learning Representations*, 2022.

[90] F. Taherkhani, A. Dabouei, S. Soleymani, J. Dawson, and N. M. Nasrabadi. Transporting labels via hierarchical optimal transport for semi-supervised learning. In *European Conference on Computer Vision*, 2020.

[91] F. Taherkhani, A. Dabouei, S. Soleymani, J. Dawson, and N. M. Nasrabadi. Self-supervised wasserstein pseudo-labeling for semi-supervised image classification. In *Conference on Computer Vision and Pattern Recognition*, 2021.

[92] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *International Conference on Computer Vision*, 2021.

[93] M. Zheng, F. Wang, S. You, C. Qian, C. Zhang, X. Wang, and C. Xu. Weakly supervised contrastive learning. In *International Conference on Computer Vision*, 2021.

[94] M. Azabou, M. G. Azar, R. Liu, C.-H. Lin, E. C. Johnson, K. Bhaskaran-Nair, M. Dabagia, B. Avila-Pires, L. Kitchell, K. B. Hengen, et al. Mine your own view: Self-supervised learning through across-sample prediction. *arXiv preprint arXiv:2102.10106*, 2021.

[95] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. In *Neural Information Processing Systems*, 2020.

[96] M. Assran, N. Ballas, L. Castrejon, and M. Rabbat. Supervision accelerates pre-training in contrastive semi-supervised learning of visual representations. *arXiv preprint arXiv:2006.10803*, 2020.

[97] B. Qu, X. Li, D. Tao, and X. Lu. Deep semantic understanding of high resolution remote sensing image. In *International conference on computer, information and telecommunication systems*, 2016.

[98] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt. Openclip, 2021.

[99] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 2016.

[100] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

A Detailed explanations of baselines and S-CLIP

A.1 Baselines

Our proposed S-CLIP includes two novel pseudo-labeling methods. Thus, we mainly compare S-CLIP with other pseudo-labeling approaches, which are extensions of semi-supervised classification techniques. Our baselines, Hard-PL and Soft-PL, extend the hard (top-1) and soft nearest neighbor approaches for pseudo-label assignment. We then utilize these methods as caption-level pseudo-labels, similar to Eq. (3). The specific forms of each method are as follows.

Hard-PL. Given an unlabeled image u_i, Hard-PL identifies the top-1 nearest labeled image[5] x_{i^*} based on their cosine similarities, calculated as $i^* = \arg\max_j (u_i \cdot x_j)$. Subsequently, the pseudo-label of u_i is determined by the corresponding caption, denoted as $q_i = y_{i^*}$. This pseudo-label serves as the target for predicting $p(y|u_i)$, and the loss is computed as $\sum_i H(p(y|u_i), q_i)$.

Soft-PL. Soft-PL calibrates the pseudo-label of Hard-PL by considering all the relations between labeled and unlabeled images, known as soft nearest neighbor [21], instead of selecting only the top-1 nearest image. For an unlabeled image u_i and a set of labeled images $\{x_i\}$, these relations are determined by applying softmax to their cosine similarities. The pseudo-label is defined as $q_i = \sigma_\tau(u_i, \{x_i\}_{i=1}^N) \in \mathbb{R}^N$, where the pseudo-label has a higher value close to one for visually similar images. We use the same softmax temperature τ as CLIP, which is used to compute the similarities between the labeled image and text embeddings. The loss is computed as $\sum_i H(p(y|u_i), q_i)$.

A.2 S-CLIP (ours)

We explain how the optimal transport of S-CLIP is computed and its relation to Soft-PL.

Sinkhorn iteration. The entropic-regularized optimal transport problem described in Eq. (2) can be efficiently solved using the Sinkhorn-Knopp [25] algorithm. For our case, the cost function $\mathrm{C} \in \mathbb{R}^{M \times N}$ is given by the negative cosine similarity: $\mathrm{C}_{ij} = 1 - u_i \cdot x_j$. The Sinkhorn-Knopp algorithm proceeds as follows: first, it computes the exponentiated cost matrix $\mathrm{K} = \exp(\frac{\mathrm{C}}{-\lambda})$, where λ denotes the scale of the entropic regularizer. Then, the algorithm iteratively updates two vectors: $\mathbf{u} \in \mathbb{R}^{M \times 1}$ and $\mathbf{v} \in \mathbb{R}^{1 \times N}$, which are initialized with uniform probability. The transportation plan Γ can be estimated using the exponentiated cost matrix K and the current vectors $\mathbf{u}$ and $\mathbf{v}$. Specifically, it is given by $\Gamma = \mathbf{u} \ \odot \mathrm{K} \ \odot \mathbf{v}$, where $\odot$ denotes element-wise multiplication.

At the initialization zero, the transportation plan Γ is obtained by normalizing the exponentiated cost matrix K since the vectors $\mathbf{u}$ and $\mathbf{v}$ are uniform. Consequently, our caption-level pseudo-label q_i is given by $q_i = \mathrm{K}_{ij} / \sum_j \mathrm{K}_{ij}$, which is equivalent to the Soft-PL $\sigma_\lambda(u_i, x_i)$ except for the softmax temperature, which is denoted as λ. This observation motivates us to set $\lambda = \tau$, matching the same temperature as CLIP, for computing the relation between embeddings. From this perspective, our proposed OT approach enhances the robustness of Soft-PL in pseudo-label assignment, especially in cases with distribution shifts between unlabeled and labeled images. Our ablation study demonstrates that increasing the number of Sinkhorn iterations improves the training process.

After initialization, the vectors $\mathbf{u}$ and $\mathbf{v}$ are updated to align with the exponentiated cost matrix K and satisfy the flow constraints $\mathbf{p}$ and $\mathbf{q}$. Specifically, given the current vectors $\mathbf{u}^{(t)}$ and $\mathbf{v}^{(t)}$, the next iteration of vectors is obtained as $\mathbf{u}^{(t+1)} = \frac{\mathbf{p}}{\mathrm{K}\mathbf{v}^{(t)}}$ and $\mathbf{v}^{(t+1)} = \frac{\mathbf{q}}{\mathrm{K}^\top \mathbf{u}^{(t+1)}}$. This iterative process converges to the optimal solution. Our ablation study demonstrates that a small number of iterations is sufficiently effective, and we employ 10 iterations in our experiments.

[5] We also tested the version directly finding the nearest text for both Hard-PL and Soft-PL. However, the similarities in visual embeddings give more robust results, as discussed in the ablation study.

B Additional related work

Optimal transport for vision-language models. Several prior works have applied optimal transport to vision-language models. However, most of them focus on transportation between an image and a caption, connecting objects and phrases [35, 36, 85–87]. This is known as weakly supervised phrase grounding [88], and it assumes a supervised setting with image-text pairs. OTTER [89] calculates transportation between images and text, which aligns with our approach. However, they also assume a supervised setting and focus on calibrating the one-hot target of CLIP. To our knowledge, our work is the first to apply optimal transport for semi-supervised CLIP training.

Optimal transport for semi-supervised learning. Some prior works have utilized optimal transport for semi-supervised learning, specifically in image classification [90, 91]. These studies demonstrate the benefits of optimal transport in pseudo-labeling class labels. However, we argue that OT plays a more critical role in vision-language models due to inherent distribution shifts between labeled and unlabeled image captions. Our experiments reveal that naive pseudo-labeling often has detrimental effects, while OT significantly improves semi-supervised CLIP training.

Partial label learning. Partial label learning (PLL) [26] relaxes the classification problem by using a candidate set as the target instead of the exact one. This involves optimizing for the best candidate within the set. Instead of determining the hard assignment of the candidate, PRODEN [75] computes soft labels for the elements in the set to smooth the optimization process. PiCO [76] further enhances this approach by incorporating contrastive learning. Our work extends the techniques of PLL to a novel problem: keyword-level pseudo-labeling for vision-language models.

Label propagation in contrastive learning. The concept of pseudo-labeling (or label propagation) is also applied in image-only contrastive learning. In addition to using the same image with different augmentations [83, 84], some studies explore the utilization of neighboring images as positive pairs [92–94]. Similar concept can be extended to supervised [95] and semi-supervised [22, 96] contrastive learning, where images sharing the same class label are considered positive pairs.

C Limitations and broader impacts

Limitations. Our pseudo-labeling approach is effective for zero-shot classification but less effective for image-text retrieval, which requires fine-grained language understanding. This is because the current pseudo-labeling approach assumes that captions in a batch can capture the semantics of unlabeled images, which may not hold true for fine-grained contexts. To address this issue, one can increase the batch size [83] or incorporate a queue of caption embeddings [84].

In addition, the current OT formulation of caption-level pseudo-labels assumes uniform sink and source constraints. However, this assumption can be refined in the case of distribution shifts. For example, techniques from robust semi-supervised learning [68, 69] can be utilized to estimate the in-domain-ness and subsequently regularize pseudo-labeling based on this estimation.

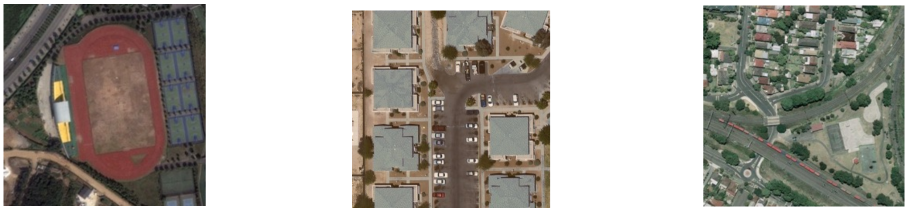
Lastly, keyword-level pseudo-labeling could be enhanced. The current method relies on the exact inclusion of keywords in the caption to form the candidate set. One can relax this criterion to include synonyms or related concepts by incorporating additional inference steps. In addition, the extracted keywords from captions may vary in their levels. Integrating a word hierarchy to account for different levels of keywords would be an interesting future research direction.

Broader impacts. Our paper aims to broaden the applicability of CLIP to specialist domains. However, this broader scope may present challenges when using the model with domains that include harmful content. These challenges arise from the data itself, rather than the model itself. Therefore, it is crucial to have adequate data regularization to effectively address these concerns.

D Experimental details

D.1 Dataset details

Remote sensing. We use the union of RSICD [27], UCM [32], and Sydney [77] datasets for training. The captions for the UCM and Sydney datasets were annotated by Qu et al. [97]. Figure 4 displays the example images and captions, while Table 6 presents the statistics of each dataset. Since an image may have multiple associated captions, we randomly select one for each iteration.



"a playground with eight basketball fields near it is surrounded by many green trees"

(a) RSICD

"many houses arranged neatly and some roads crossed them with some plants in the roadside"

(b) UCM

"a residential area with houses arranged neatly and some roads go through this area"

(c) Sydney

Figure 4: Examples of image-text pairs in the remote sensing domain.

Table 6: Statistics of image-text paired datasets in the remote sensing domain.

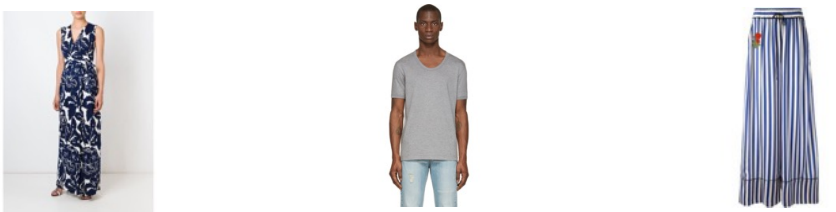
	RSICD	UCM	Sydney
# of pairs	8,734	1,680	497

We use the validation sets of RSCID-CLS [27], UCM-CLS [32], WHU-RS19 [31], RSSCN7 [79], and AID [80] datasets for zero-shot classification. RSCID-CLS and UCM-CLS are the classification versions of RSCID and UCM, respectively. Table 7 presents the statistics of each dataset.

Table 7: Statistics of classification datasets in the remote sensing domain.

	RSICD-CLS	UCM-CLS	WHU-RS19	RSSCN7	AID
# of images	1,094	2,100	1,005	2,800	10,000
# of classes	31	21	19	7	30

Fashion. We use the union of Fashion200k [28], FashionGen [81], and Polyvore Outfits [82] datasets for training. Figure 5 displays the example images and captions, while Table 8 presents the statistics of each dataset. Since an item may have multiple associated views of images corresponding to the caption, we randomly select one view for each iteration. For captions, we concatenate the titles and descriptions from FashionGen and Polyvore, and use the sentences from Fashion200k.



"white printed silk-jersey wrap dress midnight blue"

(a) Fashion200k

"grey classic u-neck t-shirt. short sleeve t-shirt in heather grey. tonal logo flag at u-neck collar."

(b) FashionGen

"off-white striped flared pants. multicoloured cotton and silk striped flared pants from off-white."

(c) Polyvore

Figure 5: Examples of image-text pairs in the fashion domain.

Table 8: Statistics of image-text paired datasets in the fashion domain.

	Fashion200k	FashionGen	Polyvore
# of pairs	61,753	60,147	71,967

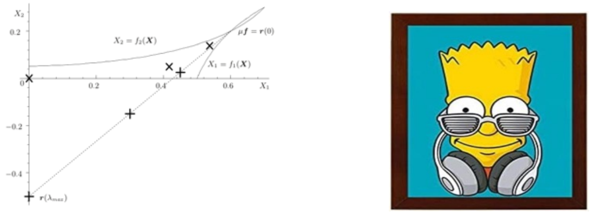
We use the validation sets of Fashion200k, FashionGen, and Polyvore datasets for zero-shot classification. Table 9 presents the statistics of each dataset. For Fashion200k and FashionGen, we report both the super-class and the sub-class. The class names in Polyvore match the same level of the super-class as those in the other datasets. For example, the super-class contains class names such as "dress," while the sub-class contains class names such as "casual and day dress."

Table 9: Statistics of classification datasets in the fashion domain.

	Fashion200k		FashionGen		Polyvore
	Super-class	Sub-class	Super-class	Sub-class	Class
# of images	29,785		32,528		14,657
# of classes	5	31	48	121	11

Scientific figures. We use the SciCap [29] dataset for training. Figure 6 displays the example images and captions, while Table 10 presents the statistics of each dataset. We use figures that do not have subfigures, denoted as the "SciCap-No-Subfig-Img" subset, for simplicity.

Comics. We use the Simpsons [30] dataset as labeled data and the Simpsons Character [33] dataset as unlabeled data. Figure 6 displays the example images and captions, while Table 10 presents the statistics of each dataset. We use 90% of pairs from the Simpsons dataset as the training split and the remaining 10% as the validation split. Since the training split contains only 720 image-text pairs, it is necessary to incorporate images from different datasets like Simpsons Characters.



"Fig. 2. variable 2-D initial probe distribution used for CFO runs reported here"

"simpsons wearing headphones and looking like a guy"

(a) SciCap

(b) Simpsons

Figure 6: Examples of image-text pairs in the scientific figure and comics domains.

Table 10: Statistics of image-text paired datasets in the scientific figure and comics domains.

	SciCap	Simpsons
# of pairs	106,834	720

D.2 Implementation details

Training. We utilize the OpenCLIP [98] library to implement our algorithms. By default, we follow the original training configuration. We use a batch size of 64 per GPU, with a total of 4 GPUs. The learning rate is set to 5e-5, and we apply the default cosine learning rate scheduling with a warmup period of the first 10 steps. We train all models until the performance saturate, which can vary over the number of image-text pairs. Specifically, for remote sensing, we train models for 25 epochs, for fashion 10 epochs, for scientific figures 5 epochs, and for comics datasets 10 epochs. We run three trials and report the average and standard deviation for all experiments.

We add the supervised CLIP loss and pseudo-label losses for semi-supervised methods. We halve the scale of pseudo-label losses to match the scale of the image part in the CLIP loss. We set the scale of the entropic regularizer to be the same as the softmax temperature of CLIP, denoted as $= \tau$, and set the number of Sinkhorn iterations to 10. Note that our method does not introduce any additional hyperparameters, except for the selection of keyword candidate sets discussed below.

Keywords. We use the available class names for remote sensing and fashion datasets while extracting the keywords using the YAKE [74] algorithm from the captions for the SciCap and Simpsons datasets. For remote sensing, we combine the class names from RSICD-CLS and UCM-CLS to obtain 45 keywords, of which around 30 words appeared in the captions. For fashion, we combine the super-class names from Fashion200k and FashionGen to obtain 56 keywords, which are more effective than sub-class names since the captions often do not contain the exact sub-class names.

Our ablation study presents that the keywords extracted by YAKE are also effective, even outperforming the class names in some cases. In the case of remote sensing, the YAKE keywords contain not only nouns like “building” but also adjectives like “green,” “many,” or “large,” which help the model understand the fine-grained information of the scene. Moreover, we utilize YAKE-100 keywords for SciCap and Simpsons when class names are unavailable, and this approach has proven to be effective. Thus, we believe that a reasonable choice of keywords would firmly benefit our method.

Computation. The computation bottleneck mostly lies in computing embeddings rather than the loss function. Thus, all methods require a similar amount of time for a single training iteration, except for S-CLIP, which takes longer due to the additional forwarding of keywords. Training duration varies depending on the number of image-text pairs, taking a few minutes to a few hours. Semi-supervised methods take twice as long as supervised CLIP since we draw batches of labeled and unlabeled data with an even size, necessitating twice as many iterations for the same epoch.

Evaluation. We use the evaluation code provided by OpenCLIP. For zero-shot classification, we use the default template: “a photo of a [class].” We use this template for fashion datasets but use the template “an aerial photograph of a [class]” for remote sensing, following the setup of [64].

E Additional remote sensing results

E.1 Different neural architectures

Table 11 presents the zero-shot classification results using the ResNet [99] and ViT [100] models. S-CLIP significantly improves the supervised baseline in all considered scenarios. Note that the larger model shows better overall accuracy, for both the original and fine-tuned CLIP. Here, S-CLIP gives reliable improvements when scaling up to larger models.

Table 11: Zero-shot classification results on remote sensing datasets using different neural architectures. Parentheses indicate the performance gap from the supervised CLIP, and bolds denote the best results within each architecture. S-CLIP give consistent improvements, scaling to larger networks.

Model	Method	RSICD-CLS	UCM-CLS	WHU-RS19	RSSCN7	AID
ResNet-50	CLIP (original)	45.3	50.5	65.5	58.9	47.8
	CLIP (fine-tune)	58.3±0.3	63.5±3.4	76.5±3.2	61.9±1.2	63.1±1.3
	S-CLIP (ours)	**66.9**±1.7 (+8.6)	**66.7**±1.6 (+3.2)	**86.9**±2.0 (+10.4)	**66.2**±1.1 (+4.3)	**73.0**±0.3 (+9.9)
ViT-B/32	CLIP (original)	55.8	58.6	76.4	62.1	55.7
	CLIP (fine-tune)	72.3±1.2	76.8±3.0	90.4±0.6	68.9±2.2	78.9±0.5
	S-CLIP (ours)	**77.5**±1.0 (+5.2)	**78.3**±3.6 (+1.5)	**93.2**±1.9 (+2.8)	**69.6**±0.8 (+0.7)	**83.7**±1.8 (+4.8)
ViT-B/16	CLIP (original)	58.9	60.1	80.9	69.4	59.6
	CLIP (fine-tune)	72.4±0.6	78.7±2.7	90.0±1.4	71.9±3.8	77.9±0.9
	S-CLIP (ours)	**79.5**±1.5 (+7.1)	**82.3**±0.4 (+3.6)	**93.9**±2.1 (+2.1)	**76.3**±0.3 (+4.4)	**85.2**±0.9 (+7.3)

E.2 Different image-text pair ratios

Table 12 presents the zero-shot classification results for different image-text pair ratios. Using only 10% of pairs (2× fewer than CLIP), S-CLIP achieves performance comparable to CLIP using 20% of pairs. Moreover, with 30% of pairs (3× fewer than CLIP), S-CLIP matches the performance of CLIP using 100% of pairs and even outperforms it, possibly due to the keyword-level pseudo-labeling.

Table 12: Zero-shot classification results on remote sensing datasets using different image-text pair ratios. Parentheses indicate the performance gap from the supervised CLIP, and bolds denote the best results within each pair ratio. S-CLIP gives consistent improvements, reducing required pairs.

Method	Ratio	RSICD-CLS	UCM-CLS	WHU-RS19	RSSCN7	AID
CLIP (original)	0%	45.3	50.5	65.5	58.9	47.8
CLIP (fine-tune)	10%	58.3±0.3	63.5±3.4	76.5±3.2	61.9±1.2	63.1±1.3
S-CLIP (ours)		**66.9**±1.7 (+8.6)	**66.7**±1.6 (+3.2)	**86.9**±2.0 (+10.4)	**66.2**±1.1 (+4.3)	**73.0**±0.3 (+9.9)
CLIP (fine-tune)	20%	66.2±1.1	69.7±0.2	82.8±1.5	65.6±3.4	71.8±1.1
S-CLIP (ours)		**77.6**±1.1 (+11.4)	**72.0**±0.7 (+2.3)	**90.7**±1.9 (+7.9)	**67.9**±2.6 (+2.3)	**83.7**±0.7 (+11.9)
CLIP (fine-tune)	30%	70.8±1.6	70.7±2.3	85.4±1.7	63.6±1.5	75.7±0.8
S-CLIP (ours)		**80.1**±1.7 (+9.3)	**74.0**±1.1 (+3.3)	**94.8**±1.2 (+9.4)	**69.6**±2.8 (+6.0)	**87.8**±0.5 (+12.1)
CLIP (fine-tune)	100%	77.5±1.5	76.5±0.5	93.9±0.9	71.2±0.3	83.6±1.6

E.3 Using all image-text pairs

We used 10% of the image-text pairs as labeled data to compare two scenarios, L=U and L≠U, in the main paper. Having demonstrated the effectiveness of S-CLIP in the L≠U scenario, the next question is whether it would also be helpful when using all paired data and extra unlabeled data. Table 13 presents the results of using the entire RS-ALL dataset as labeled data and RESISC45 as the unlabeled data. S-CLIP is shown to be effective even in this challenging scenario.

Table 13: Zero-shot classification results on remote sensing datasets using all captioned images and extra unlabeled images, where the unlabeled data is from a different distribution than the labeled data. S-CLIP also provides improvements in this most challenging scenario.

Model	Method	RSICD-CLS	UCM-CLS	WHU-RS19	RSSCN7	AID
ResNet-50	CLIP (fine-tune)	77.5±1.5	76.5±0.5	93.9±0.9	71.2±0.3	83.6±1.6
	S-CLIP (ours)	**81.1**±0.2 (+3.6)	**77.7**±1.4 (+1.2)	**95.8**±0.9 (+1.9)	**72.3**±2.8 (+1.1)	**87.7**±0.4 (+4.1)
ViT-B/32	CLIP (fine-tune)	85.3±1.0	87.3±0.7	95.8±1.0	75.7±1.4	92.0±0.5
	S-CLIP (ours)	**87.0**±0.6 (+1.7)	**88.6**±1.4 (+1.3)	**97.2**±0.4 (+1.4)	**76.2**±0.2 (+0.5)	**92.5**±0.8 (+0.5)
ViT-B/16	CLIP (fine-tune)	87.0±0.3	88.5±0.6	96.7±0.1	78.4±0.8	89.2±1.4
	S-CLIP (ours)	**87.4**±1.2 (+0.4)	**88.9**±1.6 (+0.4)	**97.3**±0.4 (+0.6)	**79.1**±1.0 (+0.7)	**93.1**±1.4 (+3.9)

E.4 Image-text retrieval R@1

Table 14 presents the R@1 results of image-text retrieval, complementing Table 2 in the main paper which shows the R@5 results. S-CLIP outperforms the semi-supervised methods in most cases, except the supervised baseline works best for image→text retrieval on the UCM dataset.

Table 14: Image-text retrieval results on remote sensing datasets, following the same setup of Table 1. Bolds denote the best results among the semi-supervised methods within the same setups. S-CLIP performs the best in most cases, similar to the observation in Table 2.

		Image→text R@1			Text→image R@1		
Method	Data	RSICD	UCM	Sydney	RSICD	UCM	Sydney
CLIP (original)	-	2.1	7.1	10.3	2.2	7.1	20.7
CLIP (fine-tune)	L	3.6±0.2	10.6±2.1	13.2±2.0	3.7±0.3	10.0±0.5	14.4±2.0
Hard-PL [16]	L=U	3.8±0.1	9.7±2.7	14.4±1.0	3.9±0.7	9.2±1.8	10.9±1.0
Soft-PL [21]		4.0±0.3	11.0±0.8	11.5±2.0	**4.2**±0.4	9.8±1.5	16.1±4.3
S-CLIP (ours)		**4.2**±0.1	**11.6**±1.1	**14.9**±1.0	**4.2**±0.6	**11.1**±0.5	**17.8**±1.0
Hard-PL [16]	L≠U	3.8±0.3	8.1±1.0	13.4±1.2	3.4±0.2	8.9±1.4	12.6±1.0
Soft-PL [21]		3.8±0.4	8.4±1.2	10.9±1.0	3.5±0.3	9.4±1.0	14.9±1.0
S-CLIP (ours)		**4.2**±0.6	**9.8**±0.6	**13.8**±0.2	**3.9**±1.1	**10.8**±1.5	**17.8**±5.0

F Additional ablation studies

F.1 Further discussions on design choices

We provide further discussions on each point of the ablation study in Section 5.4.

Training objective. The superiority of CLIP comes from the expressive information in captions. While the keyword-level pseudo-label helps the model understand specific words, it may not suffice for higher-level semantic understanding represented by a combination of words. Thus, the caption-level pseudo-label is essential for fine-grained language comprehension.

Sinkhorn iteration. Increasing the number of balances the pseudo-labels to be evenly distributed. For example, the pseudo-labels from Soft-PL often collapse to some index in the batch. This is greatly regularized in a few Sinkhorn iterations, where the maximum confidence of prediction is regularized from 100% to 40%. This regularization is more important in the L$\neq$U setup, where Soft-PL often harms the performance while ours consistently provides an improvement.

Choice of keywords. Keywords extracted by YAKE are comparable to class names. YAKE provides both nouns, such as "building," and adjectives like "green," "many," or "large," in contrast to the class names, which only contain nouns. This allows YAKE to understand the diverse semantics beyond object recognition. Thus, keyword-level pseudo-labeling is applicable to any domain.

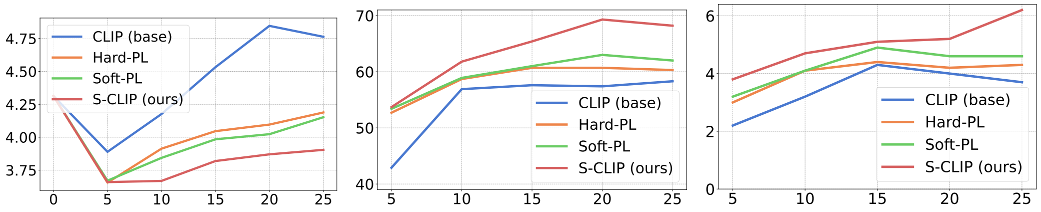
Usage of OT plan. Given the pseudo-label $q_i \in \mathbb{R}^N$ for an unlabeled image u_i, one can synthesize the pseudo-embedding $z_i = \sum_j q_{ij} \cdot y_j \in \mathbb{R}^d$, where q_{ij} denotes the j-th element of q_i, and d is the dimension of the embeddings y_j. Then, one can think of $\{u_i, z_i\}$ as paired data and apply contrastive learning to it in addition to the true pairs $\{x_i, y_i\}$. However, we observe that this pseudo-embedding is less effective than our proposed pseudo-labeling approach. This is because accurately estimating the pseudo-embedding is more challenging than assigning soft probabilities to the true embeddings. Additionally, the pseudo-embeddings lie on an interpolation of true embeddings, which can introduce confusing negatives and potentially damage contrastive learning.

Embedding for OT. We compute pseudo-labels by examining the relation between unlabeled images and either labeled images or labeled texts directly. The relation with labeled images proves to be more effective in assigning pseudo-labels. Specifically, directly using text embeddings also works reasonably well in the L=U setup. However, it often fails entirely in the L$\neq$U setup. This is because the corresponding captions of unlabeled images are unknown. In contrast, visual similarities can be robustly computed even when the images are unseen.

Partial label loss. We use the soft label given by $\sigma_\tau(u_i, \mathsf{K}_u)$ as the target for PLL. We also test the hard version, using the top-1 similar keyword from the candidate set, i.e., $i^* = \arg\max_j (u_i \cdot k_j)$. However, it performs worse, aligning with the observations from prior PLL literature.

F.2 Training curves

Figure 7 presents the training curves showing the loss and evaluation metrics. The validation accuracy for zero-shot classification on RSICD-CLS and image→text retrieval on RS-ALL is reported. The trends are similar in other datasets. Supervised CLIP is prone to overfitting, as indicated by the CLIP validation loss, but this is regularized by Hard-PL and Soft-PL. Our proposed S-CLIP yields the greatest benefits, with superior classification and retrieval performance.



(a) CLIP validation loss (b) Zero-shot classification (c) Image-text retrieval

Figure 7: Trends in loss and evaluation metrics during training. S-CLIP regularizes overfitting, as shown by the CLIP validation loss. As a result, it performs the best classification and retrieval.

F.3 Locked image or text tuning

Table 15 presents the results of fine-tuning after locking (freezing weights) either the image or text encoders, as suggested by LiT [41]. In our method, locking either encoder leads to a degradation in performance. This occurs because both the image and text data from the specialist domain are unseen, requiring the model to learn this novel information. In contrast, in the base CLIP, locking the text encoder yields similar results to having no lock.[6] This is due to the base CLIP suffering from overfitting and inadequately learning the text information. These results provide additional evidence that pseudo-labeling with S-CLIP improves the language understanding of models.

Table 15: Ablation study on locked image or text encoders. We report the average values of zero-shot accuracy (left) and image↔text retrieval (right). Locking either the image or text encoder degrades performance, as it prevents the model learning new information from specialist domains.

	Locked image		Locked text		No lock	
CLIP (fine-tune)	48.7	8.6	64.8	9.4	64.7	9.3
S-CLIP (ours)	63.1	8.7	67.1	10.2	**71.9**	**10.6**

[6] The reported average values may appear similar, but each dataset has distinct values.