Vision-Language Models for Vision Tasks: A Survey

Jingyi Zhang, *Graduate Student Member, IEEE*, Jiaxing Huang, *Graduate Student Member, IEEE*, Sheng Jin, and Shijian Lu

(Survey Paper)

***Abstract*—Most visual recognition studies rely heavily on crowd-labelled data in deep neural networks (DNNs) training, and they usually train a DNN for each single visual recognition task, leading to a laborious and time-consuming visual recognition paradigm. To address the two challenges, Vision-Language Models (VLMs) have been intensively investigated recently, which learns rich vision-language correlation from web-scale image-text pairs that are almost infinitely available on the Internet and enables zero-shot predictions on various visual recognition tasks with a single VLM. This paper provides a systematic review of visual language models for various visual recognition tasks, including: (1) the background that introduces the development of visual recognition paradigms; (2) the foundations of VLM that summarize the widely-adopted network architectures, pre-training objectives, and downstream tasks; (3) the widely-adopted datasets in VLM pre-training and evaluations; (4) the review and categorization of existing VLM pre-training methods, VLM transfer learning methods, and VLM knowledge distillation methods; (5) the benchmarking, analysis and discussion of the reviewed methods; (6) several research challenges and potential research directions that could be pursued in the future VLM studies for visual recognition.**

***Index Terms*—Big Data, big model, deep learning, deep neural network, knowledge distillation, object detection, pre-training, semantic segmentation, transfer learning, vision-language model, visual recognition, image classification.**

I. Introduction

VISUAL recognition (e.g., image classification, object detection and semantic segmentation) is a long-standing challenge in computer vision research, and it is also the cornerstone of a myriad of computer vision applications in autonomous driving [1], remote sensing [2], robotics [3], etc. With the advent of deep learning [4], [5], [6], visual recognition research has achieved great success by leveraging end-to-end trainable deep neural networks (DNNs). However, the shift from *Traditional Machine Learning* [7], [8], [9] toward deep learning comes with two new grand challenges, namely, the slow convergence of DNN training under the classical setup of *Deep Learning from Scratch* [4], [5], [6] and the laborious collection of large-scale, task-specific, and crowd-labelled data [10] in DNN training.

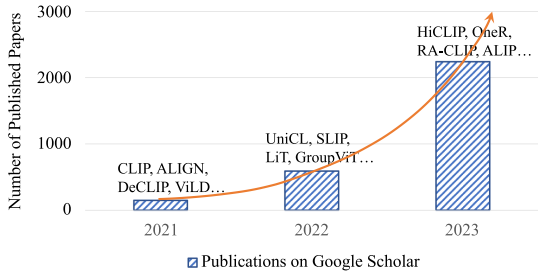

Fig. 1. Number of publications on visual recognition VLMs (from Google Scholar). The publications grow exponentially since the pioneer study CLIP [10] in 2021.

Recently, a new learning paradigm *Pre-training, Fine-tuning and Prediction* has demonstrated great effectiveness in a wide range of visual recognition tasks [11], [12], [13]. Under this new paradigm, a DNN model is first pre-trained with certain off-the-shelf large-scale training data, being annotated or unannotated, and the pre-trained model is then fine-tuned with task-specific annotated training data as illustrated in Fig. 2(a) and (b). With comprehensive knowledge learned in the pre-trained models, this learning paradigm can accelerate network convergence and train well-performing models for various downstream tasks.

Nevertheless, the *Pre-training, Fine-tuning and Prediction* paradigm still requires an additional stage of task-specific fine-tuning with labelled training data from each downstream task. Inspired by the advances in natural language processing [14], [15], [16], a new deep learning paradigm named *Vision-Language Model Pre-training and Zero-shot Prediction* has attracted increasing attention recently [10], [17], [18]. In this paradigm, a vision-language model (VLM) is pre-trained with large-scale image-text pairs that are almost infinitely available on the internet, and the pre-trained VLM can be directly applied to downstream visual recognition tasks without fine-tuning as illustrated in Fig. 2(c). The VLM pre-training is usually guided by certain vision-language objectives [10], [18], [19] that enable to learn image-text correspondences from the large-scale image-text pairs [20], [21], e.g., CLIP [10] employs an image-text contrastive objective and learns by pulling the paired images and texts close and pushing others faraway in the embedding space. In this way, the pre-trained VLMs capture rich vision-language

Manuscript received 22 April 2023; revised 1 January 2024; accepted 12 February 2024. Date of publication 26 February 2024; date of current version 2 July 2024. This work was supported in part by the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, and in part by cash and in-kind contribution from the industry partner(s). Recommended for acceptance by L. Cao. *(Jingyi Zhang and Jiaxing Huang contributed equally to this work.) (Corresponding author: Shijian Lu.)*

The authors are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: jingyi.zhang@ntu.edu.sg; shijian.lu@ntu.edu.sg).

A project associated with this survey has been created at https://github.com/jingyi0000/VLM_survey.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2024.3369699, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2024.3369699

0162-8828 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

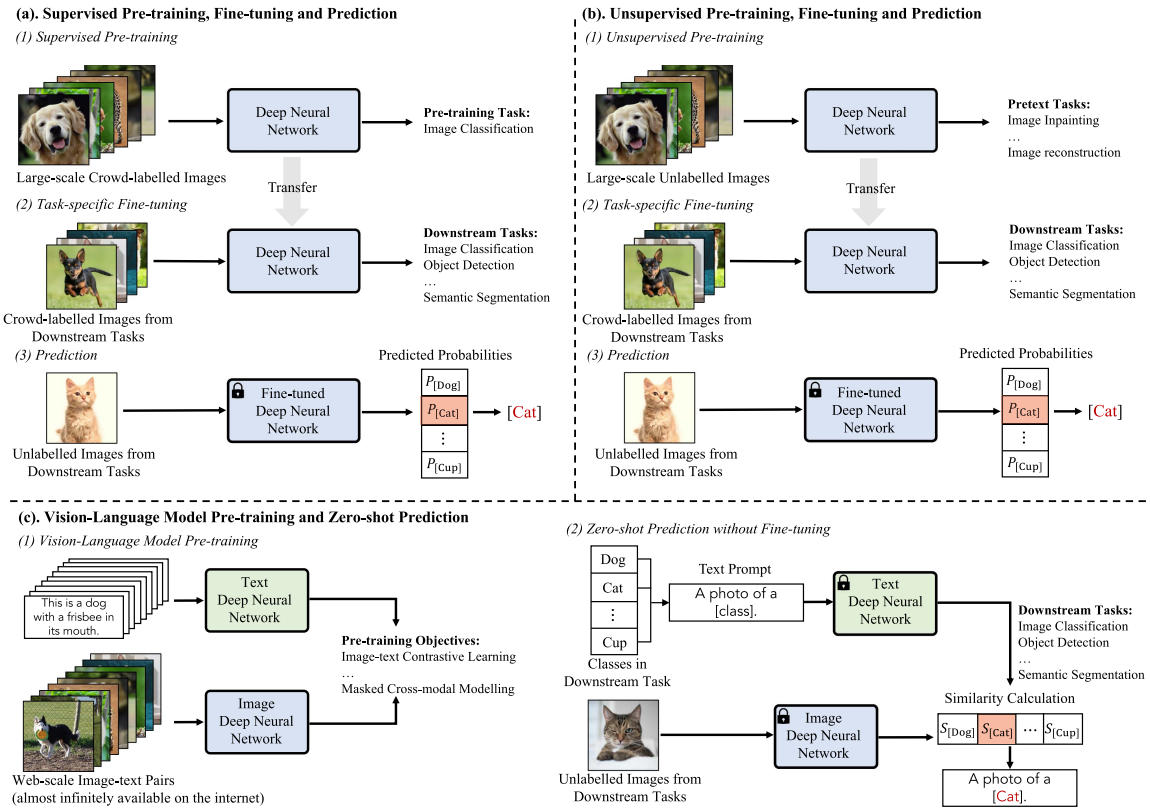

Fig. 2. Three DNN training paradigms in visual recognition. Compared with the paradigms in (a) and (b) that requires fine-tuning for each specific task with task-specific labelled data, the new learning paradigm with VLMs in (c) enables effective usage of web data and zero-shot predictions without task-specific fine-tuning.

correspondence knowledge and can perform zero-shot predictions by matching the embeddings of any given images and texts. This new learning paradigm enables effective usage of web data and allows zero-shot predictions without task-specific fine-tuning, which is simple to implement yet performs incredibly well, e.g., the pre-trained CLIP has achieved superior zero-shot performance on 36 visual recognition tasks ranging from classic image classification [22], [23], [24], [25], [26] to human action and optical character recognition [10], [27], [28], [29], [30].

Following the great success of *Vision-Language Model Pre-training and Zero-shot Prediction*, two lines of research have been intensively investigated beyond various VLM pre-training studies. The first line explores VLMs with transfer learning [31], [32], [33], [34]. It is evidenced by several transfer approaches, e.g., prompt tuning [31], [32], visual adaptation [33], [34], etc., all sharing the same target for effective adaptation of pre-trained VLMs towards various downstream tasks. The second line explores VLMs with knowledge distillation [35], [36], [37], e.g., several studies [35], [36], [37] explore how to distill knowledge from VLMs to downstream tasks, aiming for better performance in object detection, semantic segmentation, etc.

Despite the intensive interest in harvesting the vast knowledge from VLMs as evidenced by a great number of recent papers as shown in Fig. 1, the research community is short of a comprehensive survey that can help sort out existing VLM-based visual recognition studies, the facing challenges, as well as future research directions. We aim to fill up this gap by performing a systematic survey of VLM studies in various visual recognition tasks including image classification, object detection, semantic segmentation, etc. We conduct the survey from different perspectives including background, foundations, datasets, technical approaches, benchmarking, and future research directions. We believe that this survey will provide a clear big picture on what we have achieved, and we could further achieve along this emerging yet very prospective research direction.

In summary, the main contributions of this work are threefold. *First*, it presents a systematic review of VLMs for visual recognition tasks including image classification, object detection and semantic segmentation. To the best of our knowledge, this is the *first* survey of VLMs for visual recognition, which provides a big picture of this promising research filed with comprehensive summary and categorization of existing studies. *Second*, it studies the up-to-date progress of VLMs for visual recognition, including a comprehensive benchmarking and discussion of existing work over multiple public datasets. *Third*, it shares several research challenges and potential research directions that could be pursued in VLMs for visual recognition.

The rest of this survey is organized as follows. Section II introduces the paradigm development of visual recognition and several related surveys. Section III describes the foundations of VLMs, including widely used deep network architectures, pre-training objectives, pre-training frameworks and downstream tasks in VLM evaluations. Section IV introduces the commonly used datasets in VLM pre-training and evaluations. Section V reviews and categorizes VLM pre-training methods. Sections VI and VII provide a systematic review of transfer learning

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

and knowledge distillation approaches for VLMs , respectively. Section VIII benchmarks the reviewed methods on multiple widely-adopted datasets. Finally, we share several promising VLM research directions in Section IX.

II. Background

This section first presents the development of the training paradigm of visual recognition and how it evolves towards the paradigm *Vision-Language Model Pre-training and Zero-shot Prediction*. Then, we introduce the development of the vision-language models (VLMs) for visual recognition. We also discuss several related surveys to highlight the scope and contributions of this survey.

A. Training Paradigms for Visual Recognition

The development of visual recognition paradigms can be broadly divided into five stages, including (1) *Traditional Machine Learning and Prediction*, (2) *Deep Learning from Scratch and Prediction*, (3) *Supervised Pre-training, Fine-tuning and Prediction*, (4) *Unsupervised Pre-training, Fine-tuning and Prediction* and (5) *Vision-language Model Pre-training and Zero-shot Prediction*. In what following, we introduce, compare and analyze the five training paradigms in detail.

1) Traditional Machine Learning and Prediction: Before the deep learning era [4], visual recognition studies rely heavily on *feature engineering* with hand-crafted features [9], [38] and lightweight learning models [7], [8], [39] that classify the hand-crafted features into pre-defined semantic categories. However, this paradigm requires domain experts for crafting effective features for specific visual recognition tasks, which does not cope with complex tasks well and also has poor scalability.

2) Deep Learning From Scratch and Prediction: With the advent of deep learning [4], [5], [6], visual recognition research has achieved great success by leveraging end-to-end trainable DNNs that circumvent the complicated *feature engineering* and allow focusing on the *architecture engineering* of neural networks for learning effective features. For example, ResNet [6] enables very deep networks by a skip design and allows learning from massive crowd-labelled data with unprecedented performance on the challenging ImageNet benchmark [40]. However, the turn from traditional machine learning toward deep learning raises two new grand challenges: the slow convergence of DNN training under the classical setup of *Deep Learning from Scratch* and the laborious collection of large-scale, task-specific, and crowd-labelled data [10] in DNN training.

3) Supervised Pre-Training, Fine-Tuning and Prediction: With the discovery that features learned from labelled large-scale datasets can be transferred to downstream tasks [11], the paradigm *Deep Learning from Scratch and Prediction* has been gradually replaced by a new paradigm of *Supervised Pre-training, Fine-tuning and Prediction*. This new learning paradigm, as illustrated in Fig. 2(a), pre-trains DNNs on large-scale labelled data (e.g., ImageNet) with a supervised loss and then fine-tunes the pre-trained DNN with task-specific training data [11]. As the pre-trained DNNs have learned certain visual knowledge, it can accelerate network convergence and help train well-performing models with limited task-specific training data.

4) Unsupervised Pre-Training, Fine-Tuning & Prediction: Though *Supervised Pre-training, Fine-tuning and Prediction* achieves state-of-the-art performance on many visual recognition tasks, it requires large-scale labelled data in pre-training. To mitigate this constraint, [12], [13] adopt a new learning paradigm *Unsupervised Pre-training, Fine-tuning and Prediction* that explores self-supervised learning to learn useful and transferable representations from unlabelled data, as illustrated in Fig. 2(b). To this end, various self-supervised training objectives [12], [41] have been proposed including masked image modelling that models cross-patch relations [41], contrastive learning that learns discriminative features by contrasting training samples [12], etc. The self-supervised pre-trained models are then fine-tuned on downstream tasks with labelled task-specific training data. Since this paradigm does not require labelled data in pre-training, it can exploit more training data for learning useful and transferable features, leading to even better performance as compared with the supervised pre-training [12], [13].

5) VLM Pre-Training and Zero-Shot Prediction: Though *Pre-training and Fine-tuning* with either supervised or unsupervised pre-training improves the network convergence, it still requires a fine-tuning stage with labelled task data as shown in Figs. 2(a) and (b). Motivated by great success in natural language processing [14], [15], [16], a new deep learning paradigm named *Vision-Language Model Pre-training and Zero-shot Prediction* has been proposed for visual recognition, as shown in Fig. 2(c). With large-scale image-text pairs that are almost infinitely available on the internet, a VLM is pre-trained by certain vision-language objectives [10], [18], [19] which captures rich vision-language knowledge and can perform zero-shot predictions (without fine-tuning) on downstream visual recognition tasks by matching the embeddings of any given images and texts.

Compared with *Pre-training and Fine-tuning*, this new paradigm enables effective use of large-scale web data and zero-shot predictions without task-specific fine-tuning. Most existing research attempts to improve VLMs from 3 perspectives: 1) collecting large-scale informative image-text data, 2) designing high-capacity models for effective learning from Big Data, 3) designing new pre-training objectives for learning effective VLMs. In this paper, we provide a systematic survey of this new vision-language learning paradigm aiming to provide a clear big picture on exiting VLM studies, the facing challenges and future directions for this challenging but promising research filed.

B. Development of VLMs for Visual Recognition

Visual recognition related VLM studies have made great progresses since the development of CLIP [10]. We present VLMs for visual recognition from three aspects as illustrated in Fig. 3: (1) *Pre-training objectives: from "a single objective" to "multiple hybrid objectives"*. Early VLMs [10], [17] generally adopt a single pre-training objective, whereas recent VLMs [18], [42] introduce multiple objectives (e.g., contrastive, alignment and generative objectives) for exploring their synergy for more robust VLMs and better performance in downstream tasks; (2) *Pre-training frameworks: from "multiple separate networks" to "a unified network"*. Early VLMs [10], [17] employ two-tower pre-training frameworks, whereas recent VLMs [43], [44]

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

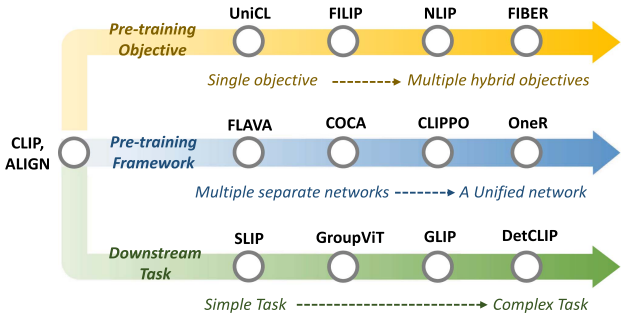

Fig. 3. Illustration of development of VLMs for visual recognition.

attempt one-tower pre-training framework that encodes images and texts with a unified network with less GPU memory usage yet more efficient communications across data modalities; 3) *Downstream tasks: from simple to complex tasks.* Early VLMs [10], [17] focus on image-level visual recognition tasks, whereas recent VLMs [45], [46] are more general-purpose which can also work for dense prediction tasks that are complex and require localization related knowledge.

C. Relevant Surveys

To the best of our knowledge, this is the *first* survey that reviews VLMs for various visual recognition tasks. Several relevant surveys have been conducted which review VLMs for vision-language tasks instead such as visual question answering [47], natural language for visual reasoning [48] , and phrase grounding [49]. For instance, Li et al. [50] shared advances on vision-language tasks, including VLM pre-training for various task-specific methods. Du et al. [51] and Chen et al. [52] reviewed VLM pre-training for vision-language tasks [47], [48], [49]. Xu et al. [53] and Wang et al. [54] shared recent progress of multi-modal learning on multi-modal tasks. Differently, as shown in Fig. 4, we review VLMs for visual recognition tasks from three major aspects: 1) Recent progress of VLM pre-training for visual recognition tasks; 2) Two typical transfer approaches from VLMs to visual recognition tasks; 3) Benchmarking of VLM pre-training methods on visual recognition tasks.

III. VLM Foundations

VLM pre-training [10], [17] aims to pre-train a VLM to learn image-text correlation, targeting effective zero-shot predictions on visual recognition tasks [6], [55], [56]. Given image-text pairs [20], [21], it first employs a text encoder and an image encoder to extract image and text features [6], [14], [57], [58] and then learns the vision-language correlation with certain pre-training objectives [10], [17]. Hence, VLMs can be evaluated on unseen data in a zero-shot manner [10], [17] by matching the embeddings of any given images and texts. This section introduces the foundations of VLM pre-training, including common network architectures for extracting image and text features, pre-training objectives for modelling vision-language correlation, frameworks for VLM pre-training and downstream tasks for VLM evaluations.

A. Network Architectures

VLM pre-training works with a deep neural network that extracts image and text features from N image-text pairs within a pre-training dataset $\mathcal{D} = \{x_n^I, x_n^T\}_{n=1}^N$, where x_n^I and x_n^T denote an image sample and its paired text sample. The deep neural network has an image encoder f_θ and a text encoder f_ϕ, which encode the image and text (from an image-text pair $\{x_n^I, x_n^T\}$) into an image embedding $z_n^I = f_\theta(x_n^I)$ and a text embedding $z_n^T = f_\phi(x_n^T)$, respectively. This section presents the architecture of widely-adopted deep neural networks in VLM pre-training.

1) Architectures for Learning Image Features: Two types of network architectures have been widely adopted to learn image features, namely, CNN-based architectures and Transformer-based architectures.

CNN-based Architectures: Different ConvNets (e.g., VGG [5], ResNet [6] and EfficientNet [59]) have been designed for learning image features. Being one of the most popular ConvNet in VLM pre-training, ResNet [6] adopts skip connections between convolution blocks which mitigates gradient vanishing and explosion and enables very deep neural networks. For better feature extraction and vision-language modelling, several studies [10] modify the original network architecture [6], [59]. Take ResNet as an example. They introduce the ResNet-D [60], employ the antialiased rect-2 blur pooling in [61], and replace the global average pooling with an attention pooling in the transformer multi-head attention [58].

Transformer-based Architectures. Transformers have recently been extensively explored in visual recognition tasks, such as image classification [57], object detection [62] and semantic segmentation [63]. As a standard Transformer architecture for image feature learning, ViT [57] employs a stack of Transformer blocks each of which consists of a multi-head self-attention layer and a feed-forward network. The input image is first split into fixed-size patches and then fed to the Transformer encoder after linear projection and position embedding. [10], [18], [64] modify ViT by adding a normalization layer before the transformer encoder.

2) Architectures for Learning Language Features: Transformer & its variants [14], [16], [58] have been widely adopted for learning text features. The standard Transformer [58] has an encoder-decoder structure, where the encoder has 6 blocks each of which has a multi-head self-attention layer and a multi-layer perceptron (MLP). The decoder also has 6 blocks each of which has a multi-head attention layer, a masked multi-head layer and a MLP. Most VLM studies such as CLIP [10] adopt the standard Transformer [58] with minor modifications as in GPT_2 [16], and train from scratch without initialization with GPT_2 weights.

B. VLM Pre-Training Objectives

As the core of VLM, various vision-language pre-training objectives [10], [12], [14], [19], [42], [65], [66], [67] have been designed for learning rich vision-language correlation. They fall broadly into three categories: contrastive objectives, generative objectives and alignment objectives.

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

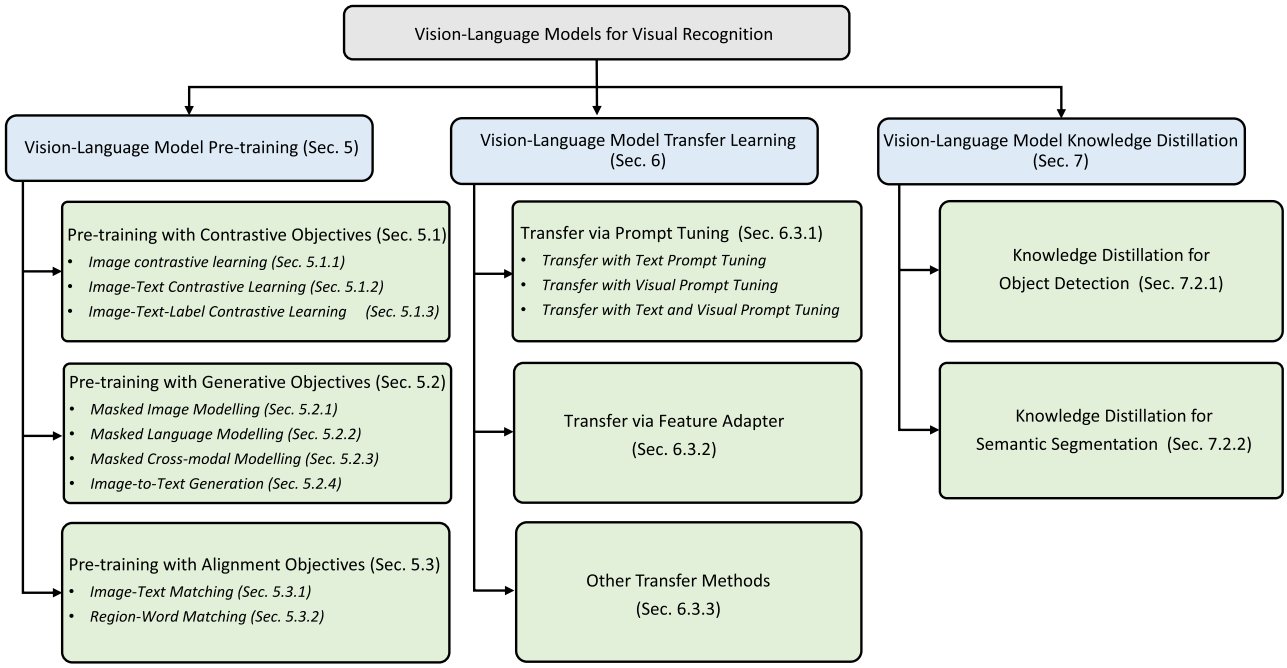

Fig. 4. Typology of vision-language models for visual recognition.

1) Contrastive Objectives: Contrastive objectives train VLMs to learn discriminative representations by pulling paired samples close and pushing others faraway in the feature space [10], [12], [65].

Image Contrastive Learning aims to learn discriminative image features [12], [13] by forcing a query image to be close with its positive keys (i.e., its data augmentations) and faraway from its negative keys (i.e., other images) in the embedding space. Given a batch of B images, contrastive-learning objectives (e.g., InfoNCE [68] and its variants [12], [13]) are usually formulated as follows:

$$\mathcal{L}_I^{\text{InfoNCE}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(z_i^I \cdot z_+^I/\tau\right)}{\sum_{j=1,j\neq i}^{B+1}\exp(z_i^I \cdot z_j^I/\tau)}, \quad (1)$$

where z_i^I is the query embedding, $\{z_j^I\}_{j=1,j\neq i}^{B+1}$ are key embeddings, where z_+^I stands for z_i^I's positive key and the rest are z_i^I's negative keys. τ is a temperature hyper-parameter that controls the density of the learned representation.

Image-Text Contrastive Learning aims to learn discriminative image-text representations by pulling the embeddings of paired images and texts close while pushing others [10], [17] away. It is usually achieved by minimizing a symmetrical image-text infoNCE loss [10], i.e., $\mathcal{L}_{\text{infoNCE}}^{IT} = \mathcal{L}_{I\to T} + \mathcal{L}_{T\to I}$, where $\mathcal{L}_{I\to T}$ contrasts the query image with the text keys while $\mathcal{L}_{T\to I}$ contrasts the query text with image keys. Given a batch of B image-text pairs, $\mathcal{L}_{I\to T}$ and $\mathcal{L}_{T\to I}$ are defined as follows:

$$\mathcal{L}_{I\to T} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(z_i^I \cdot z_i^T/\tau\right)}{\sum_{j=1}^{B}\exp(z_i^I \cdot z_j^T/\tau)}, \quad (2)$$

$$\mathcal{L}_{T\to I} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(z_i^T \cdot z_i^I/\tau\right)}{\sum_{j=1}^{B}\exp(z_i^T \cdot z_j^I/\tau)}, \quad (3)$$

where z^I and z^T stand for the image embeddings and text embeddings, respectively.

Image-Text-Label Contrastive Learning: Image-text-label contrastive learning [65] introduces Supervised Contrastive Learning [69] into image-text contrastive learning, which is defined by reformulating (2) and (3) as follows:

$$\mathcal{L}_{I\to T}^{ITL} = -\sum_{i=1}^{B}\frac{1}{|\mathcal{P}(i)|}\sum_{k\in\mathcal{P}(i)}\log\frac{\exp\left(z_i^I \cdot z_k^T/\tau\right)}{\sum_{j=1}^{B}\exp(z_i^I \cdot z_j^T/\tau)}, \quad (4)$$

$$\mathcal{L}_{T\to I}^{ITL} = -\sum_{i=1}^{B}\frac{1}{|\mathcal{P}(i)|}\sum_{k\in\mathcal{P}(i)}\log\frac{\exp\left(z_i^T \cdot z_k^I/\tau\right)}{\sum_{j=1}^{B}\exp(z_i^T \cdot z_j^I/\tau)}, \quad (5)$$

where $k \in \mathcal{P}(i) = \{k | k \in B, y_k = y_i\}$ [65] and y is the category label of (z^I, z^T). With (4) and (5), the image-text-label infoNCE loss is defined as: $\mathcal{L}_{\text{infoNCE}}^{ITL} = \mathcal{L}_{I\to T}^{ITL} + \mathcal{L}_{T\to I}^{ITL}$.

2) Generative Objectives: Generative objectives learn semantic features by training networks to generate image/text data via image generation [12], [70], language generation [14], [19], or cross-modal generation [42].

Masked Image Modelling learns cross-patch correlation by masking and reconstructing images [41], [70]. It masks a set of patches of an input image randomly and trains the encoder to reconstruct the masked patches conditioned on unmasked patches. Given a batch of B images, the loss function can be formulated as:

$$\mathcal{L}_{MIM} = -\frac{1}{B}\sum_{i=1}^{B}\log f_\theta(\overline{x}_i^I \mid \hat{x}_i^I), \quad (6)$$

where $\overline{x}_i^I$ and $\hat{x}_i^I$ denote the masked patches and the unmasked patches in x_i^I, respectively.

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

Masked Language Modelling is a widely adopted pre-training objective in NLP [14]. It randomly masks a certain percentage (e.g., 15% in BERT [14]) of the input text tokens, and reconstruct them with unmasked tokens:

$$\mathcal{L}_{MLM} = -\frac{1}{B}\sum_{i=1}^{B}\log f_\phi(\overline{x}_i^T \mid \hat{x}_i^T), \tag{7}$$

where $\overline{x}_i^T$ and $\hat{x}_i^T$ denote the masked and unmasked tokens in x_i^T, respectively. B denotes the batch size.

Masked Cross-Modal Modelling integrates masked image modelling and masked language modelling [42]. Given an image-text pair, it randomly masks a subset of image patches and a subset of text tokens and then learns to reconstruct them conditioned on unmasked image patches and unmasked text tokens as follows:

$$\mathcal{L}_{MCM} = -\frac{1}{B}\sum_{i=1}^{B}[\log f_\theta(\overline{x}_i^I|\hat{x}_i^I,\hat{x}_i^T) + \log f_\phi(\overline{x}_i^T|\hat{x}_i^I,\hat{x}_i^T)], \tag{8}$$

where $\overline{x}_i^I/\hat{x}_i^I$ denotes the masked/unmasked patches in x_i^I, $\overline{x}_i^T/\hat{x}_i^T$ denotes the masked/unmasked text tokens in x_i^T.

Image-to-Text Generation aims to predict text x^T autoregressively based on the image paired with x^T [19]:

$$\mathcal{L}_{ITG} = -\sum_{l=1}^{L}\log\ f_\theta(x^T \mid x_{<l}^T, z^I), \tag{9}$$

where L denotes the number of tokens to be predicted for x^T and z^I is the embedding of the image paired with x^T.

3) Alignment Objectives: Alignment objectives align the image-text pair via global image-text matching [71], [72] or local region-word matching [45], [67] on embedding space.

Image-Text Matching models global correlation between images and texts [71], [72], which can be formulated with a score function $\mathcal{S}(\cdot)$ that measures the alignment probability between the image and text and a binary classification loss:

$$\mathcal{L}_{IT} = p\log\mathcal{S}(z^I,z^T) + (1-p)\log(1-\mathcal{S}(z^I,z^T)), \tag{10}$$

where p is 1 if the image and text are paired and 0 otherwise.

Region-Word Matching aims to model local cross-modal correlation (i.e., between "image regions" and "words") in image-text pairs [45], [67] for dense visual recognition tasks such as object detection. It can be formulated as:

$$\mathcal{L}_{RW} = p\log\mathcal{S}^r(r^I,w^T) + (1-p)\log(1-\mathcal{S}^r(r^I,w^T)), \tag{11}$$

where (r^I, w^T) denotes a region-word pair and $p = 1$ if the region and word are paired otherwise $p = 0$. $\mathcal{S}^r(\cdot)$ denotes a local score function that measures the similarity between "image regions" and "words".

C. VLM Pre-Training Frameworks

This section presents widely adopted frameworks in VLM pre-training, including two-tower, two-leg and one-tower pre-training frameworks.

Specifically, two-tower framework has been widely adopted in VLM pre-training [10], [17], where input images and texts are encoded with two separate encoders respectively, as illustrated in Fig. 5(a). Slightly differently, two-leg framework [19], [42] introduces additional multi-modal fusion layers which enable feature interaction between image and text modalities, as illustrated in Fig. 5(b). As a comparison, one-tower VLMs [43], [44] attempt to unify vision and language learning in a single encoder as illustrated in Fig. 5(c), aiming to facilitate efficient communications across data modalities.

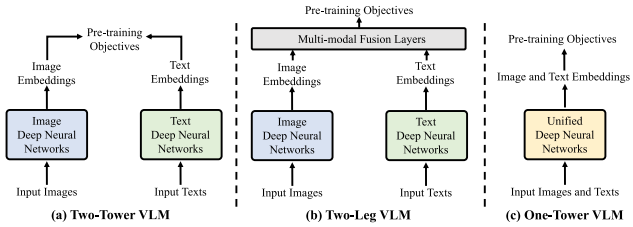

Fig. 5. Illustration of typical VLM pre-training frameworks.

D. Evaluation Setups and Downstream Tasks

This section presents widely adopted setups and downstream tasks in VLM evaluation. The setups include *zero-shot prediction* and *linear probing*, and the downstream tasks include image classification, object detection, semantic segmentation, image-text retrieval, and action recognition.

1) Zero-Shot Prediction: As the most common way of evaluating VLMs' generalization capability [10], [17], [18], [64], [84], zero-shot prediction directly applies pre-trained VLMs to downstream tasks without any task-specific fine-tuning [10].

Image Classification [5], [6] aims to classify images into pre-defined categories. VLMs achieve zero-shot image classification by comparing the embeddings of images and texts, where "prompt engineering" is often employed to generate task-related prompts like "a photo of a `[label]`." [10].

Semantic Segmentation [56] aims to assign a category label to each pixel in images. Pre-trained VLMs achieve zero-shot prediction for segmentation tasks by comparing the embeddings of the given image pixels and texts.

Object Detection [11], [55] aims to localize and classify objects in images, which is important for various vision applications. With the object locating ability learned from auxiliary datasets [85], [86], pre-trained VLMs achieve zero-shot prediction for object detection tasks by comparing the embeddings of the given object proposals and texts.

Image-Text Retrieval [87] aims to retrieve the demanded samples from one modality given the cues from another modality, which consists of two tasks, i.e., text-to-image retrieval that retrieves images based on texts and image-to-text retrieval that retrieves texts based on images.

2) Linear Probing: Linear probing has been widely adopted in VLM evaluations [10]. It freezes the pre-trained VLM and trains a linear classifier to classify the VLM-encoded embeddings to assess the VLM representations. Image classification [5], [6] and action recognition [28], [29] have been widely adopted in such evaluations, where video clips are often subsampled for efficient recognition in action recognition tasks [10].

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

TABLE I
SUMMARY OF THE WIDELY USED IMAGE-TEXT DATASETS FOR VLM PRE-TRAINING

Dataset	Year	Num. of Image-Text Pairs	Language	Public
SBU Caption [73] [link]	2011	1M	English	✓
COCO Caption [74] [link]	2016	1.5M	English	✓
Yahoo Flickr Creative Commons 100 Million (YFCC100M) [75] [link]	2016	100M	English	✓
Visual Genome (VG) [76] [link]	2017	5.4 M	English	✓
Conceptual Captions (CC3M) [77] [link]	2018	3.3M	English	✓
Localized Narratives (LN) [78] [link]	2020	0.87M	English	✓
Conceptual 12M (CC12M) [79] [link]	2021	12M	English	✓
Wikipedia-based Image Text (WIT) [80] [link]	2021	37.6M	108 Languages	✓
Red Caps (RC) [81] [link]	2021	12M	English	✓
LAION400M [21] [link]	2021	400M	English	✓
LAION5B [20] [link]	2022	5B	Over 100 Languages	✓
WuKong [82] [link]	2022	100M	Chinese	✓
CLIP [10]	2021	400M	English	✗
ALIGN [17]	2021	1.8B	English	✗
FILIP [18]	2021	300M	English	✗
WebLI [83]	2022	12B	109 Languages	✗

[link] directs to dataset websites.

TABLE II
SUMMARY OF THE WIDELY-USED VISUAL RECOGNITION DATASETS FOR VLM EVALUATION

Task	Dataset	Year	Classes	Training	Testing	Evaluation Metric
Image Classification	MNIST [88] [link]	1998	10	60,000	10,000	Accuracy
	Caltech-101 [89] [link]	2004	102	3,060	6,085	Mean Per Class
	PASCAL VOC 2007 Classification [90] [link]	2007	20	5,011	4,952	11-point mAP
	Oxford 102 Folwers [91] [link]	2008	102	2,040	6,149	Mean Per Class
	CIFAR-10 [23] [link]	2009	10	50,000	10,000	Accuracy
	CIFAR-100 [23] [link]	2009	100	50,000	10,000	Accuracy
	ImageNet-1k [40] [link]	2009	1000	1,281,167	50,000	Accuracy
	SUN397 [24] [link]	2010	397	19,850	19,850	Accuracy
	SVHN [92] [link]	2011	10	73,257	26,032	Accuracy
	STL-10 [93] [link]	2011	10	1,000	8,000	Accuracy
	GTSRB [94] [link]	2011	43	26,640	12,630	Accuracy
	KITTI Distance [1] [link]	2012	4	6,770	711	Accuracy
	IIIT5k [95] [link]	2012	36	2,000	3,000	Accuracy
	Oxford-IIIT PETS [26] [link]	2012	37	3,680	3,669	Mean Per Class
	Stanford Cars [25] [link]	2013	196	8,144	8,041	Accuracy
	FGVC Aircraft [96] [link]	2013	100	6,667	3,333	Mean Per Class
	Facial Emotion Recognition 2013 [97] [link]	2013	8	32,140	3,574	Accuracy
	Rendered SST2 [98] [link]	2013	2	7,792	1,821	Accuracy
	Describable Textures (DTD) [99] [link]	2014	47	3,760	1,880	Accuracy
	Food-101 [22] [link]	2014	102	75,750	25,250	Accuracy
	Birdsnap [100] [link]	2014	500	42,283	2,149	Accuracy
	RESISC45 [101] [link]	2017	45	3,150	25,200	Accuracy
	CLEVR Counts [102] [link]	2017	8	2,000	500	Accuracy
	PatchCamelyon [103] [link]	2018	2	294,912	32,768	Accuracy
	EuroSAT [104] [link]	2019	10	10,000	5,000	Accuracy
	Hateful Memes [27] [link]	2020	2	8,500	500	ROC AUC
	Country211 [10] [link]	2021	211	43,200	21,100	Accuracy
Image-Text Retrieval	Flickr30k [105] [link]	2014	-	31,783	-	Recall
	COCO Caption [74] [link]	2015	-	82,783	5,000	Recall
Action Recognition	UCF101 [29] [link]	2012	101	9,537	1,794	Accuracy
	Kinetics700 [30] [link]	2019	700	494,801	31,669	Mean(top1, top5)
	RareAct [28] [link]	2020	122	7,607	-	mWAP, mSAP
Object Detection	COCO 2014 Detection [106] [link]	2014	80	83,000	41,000	box mAP
	COCO 2017 Detection [106] [link]	2017	80	118,000	5,000	box mAP
	LVIS [107] [link]	2019	1203	118,000	5,000	box mAP
	ODinW [108] [link]	2022	314	132413	20070	box mAP
Semantic Segmentation	PASCAL VOC 2012 Segmentation [90] [link]	2012	20	1464	1449	mIoU
	PASCAL Content [109] [link]	2014	459	4998	5105	mIoU
	Cityscapes [110] [link]	2016	19	2975	500	mIoU
	ADE20k [111] [link]	2017	150	25574	2000	mIoU

[link] directs to dataset websites.

IV. DATASETS

This section summarizes the commonly used datasets for VLM pre-training and evaluations, as detailed in Tables I and II.

A. Datasets for Pre-Training VLMs

For VLM pre-training, multiple large-scale image-text datasets [10], [17], [20], [21] were collected from the internet. Compared with traditional crowd-labelled datasets [40], [90],

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

TABLE III
SUMMARY OF VISION-LANGUAGE MODEL PRE-TRAINING METHODS

Method	Dataset	Objective	Contribution
CLIP† [10] [code]	CLIP∗	Con	Propose image-text contrastive learning for VLM pre-training.
ALIGN† [17]	ALIGN∗	Con	Leverage large-scale noisy data to scale-up VLM pre-training data.
OTTER† [112] [code]	CC3M, YFCC15M, WIT	Con	Employ optimal transport for data efficient VLM pre-training.
DeCLIP† [113] [code]	CC3M, CC12M, YFCC100M, WIT∗	Con,Gen	Employ image/text self-supervision for data efficient VLM pre-training.
ZeroVL† [114] [code]	SBU, VG, CC3M, CC12M	Con	Introduce data augmentation for data-efficient VLM pre-training.
FILIP† [18]	FILIP∗, CC3M, CC12M, YFCC100M	Con,Align	Leverage region-word similarity for fine-grained VLM pre-training.
UniCL† [65] [code]	CC3M, CC12M, YFCC100M	Con	Propose image-text-label contrastive learning for VLM pre-training.
Florence† [115]	FLD-900M∗	Con	Scale up pre-training data and include depth and temporal information.
SLIP† [64] [code]	YFCC100M	Con	Introduce image self-supervision learning into VLM pre-training.
PyramidCLIP† [116]	SBU, CC3M, CC12M, YFCC100M, LAION400M	Con	Perform peer-level/cross-level contrastive learning within/across multiple semantic levels.
ChineseCLIP† [117] [code]	LAION5B, WuKong, VG, COCO	Con	Collect large-scale Chinese image-text data and Introduce Chinese VLM.
LiT† [118] [project]	CC12M, YFCC100M, WIT∗	Con	Propose contrastive tuning with the locked image encoder.
AltCLIP† [119] [code]	WuDao, LAION2B, LAION5B	Con	Leverage the multilingual text encoder to achieve multilingual VLM.
FLAVA‡ [42] [code]	COCO, SBU, LN, CC3M, VG, WIT, CC12M, RC, YFCC100M	Gen,Con,Align	Propose a universal and foundational VLM that tackles the single-modal (*i.e.*, image or text) and the multi-model cases at the same time.
KELIP† [120] [code]	CUB200, WIT, YFCC15M, CC3M, CC12M, LAION400M, K-WIT∗	Con,Gen	Collect large-scale Korean image-text pair data and develop bilingual VLMs with Korean and English.
COCA‡ [19] [code]	ALIGN∗	Con,Gen	Combine contrastive learning and image captioning for pre-training.
nCLIP† [121]	COCO, VG, SBU, CC3M, CC12M, YFCC14M	Con,Align	Propose a non-contrastive pre-training objective (*i.e.*, a cross-entropy loss for global image-text matching) for VLM pre-training.
K-lite† [122] [code]	CC3M, CC12M, YFCC100M	Con	Leverage auxiliary datasets for training transferable VLMs.
NLIP‡ [123]	YFCC100M, COCO	Con,Gen	Train noise-robust VLM via noise harmonization and completion.
UniCLIP† [84]	CC3M, CC12M, YMCC100M	Con	Propose unified image-text and image-image contrastive learning.
PaLI‡ [83] [project]	WebLI∗	Gen	Scale up the data, model and language in VLM pre-taring.
HiCLIP† [124] [code]	YFCC100M, CC3M, CC12M	Con	Propose to incorporate hierarchy-aware attention into VLM pre-training.
CLIPPO§ [43] [code]	WebLI∗	Con	Learn image and text data with a single network for VLM pre-training.
OneR§ [44]	CC3M, SBU, VG, COCO	Con,Gen	Unify image and text learning in a single tower transformer.
RA-CLIP† [125]	YFCC100M	Con	Propose retrieval-augmented image-text contrastive learning.
LA-CLIP† [126] [code]	CC3M, CC12M, RC, LAION400M	Con	Propose LLMs-augmented image-text contrastive learning.
ALIP† [127] [code]	YFCC100M	Con	Introduce synthetic caption supervision into VLM pre-training.
GrowCLIP‡ [128]	CC12M	Con	Propose online-learning image-text contrastive learning.
GroupVit† [129] [code]	CC12M, YMCC100M	Con	Propose hierarchical visual concepts grouping for VLM pre-training.
SegClip† [46] [code]	CC3M, COCO	Con,Gen	Propose a plug-in semantic group module for VLM pre-training.
CLIPpy† [130] [code]	CC12M	Con	Propose spatial representation aggregation for VLM pre-training.
RegionClip† [131] [code]	CC3M, COCO	Con,Align	Learn region-level visual representations for VLM pre-training.
GLIP‡ [67] [code]	CC3M, CC12M, SBU	Align	Unify detection and phrase grounding for grounded VLM pre-training.
FIBER‡ [71] [code]	COCO, CC3M, SBU, VG	Con,Gen,Align	Propose deep multi-modal fusion for coarse-to-fine VLM pre-training.
DetCLIP‡ [45]	YMCC100M	Align	Present a paralleled visual-concept VLM pre-training method.

Con: Contrastive Objective; Gen: Generative Objective; Align: Alignment Objective. †, ‡ and § denote two-tower, two-leg and one-tower pre-training frameworks, respectively. *denotes non-public datasets. [code] directs to code websites.

[110], the image-text datasets [10], [21] are much larger and cheaper to collect. For example, recent image-text datasets are generally at billion scale [20], [21], [83]. Beyond image-text datasets, several studies [19], [43], [45], [67] utilize auxiliary datasets to provide additional information for better vision-language modelling, e.g., GLIP [67] leverages Object365 [85] for extracting region-level features. The details of image-text datasets and auxiliary datasets for VLM pre-training are provided in Appendix B, available online.

B. Datasets for VLM Evaluation

Many datasets have been adopted in VLM evaluations as shown in Table II, including 27 for image classification, 4 for object detection, 4 for semantic segmentation, 2 for image-text retrieval, and 3 for action recognition (dataset details provided in Appendix C, available online). For example, the 27 image classification datasets cover a wide range of visual recognition tasks from fine-grained tasks like Oxford-IIIT PETS [26] for pet identification and Stanford Cars [25] for car recognition, to general tasks like ImageNet [40].

V. VISION-LANGUAGE MODEL PRE-TRAINING

VLM pre-training has been explored with three typical objectives: contrastive objectives, generative objectives and alignment objectives. This section reviews them with multiple VLM pre-training studies as listed in Table III.

A. VLM Pre-Training With Contrastive Objectives

Contrastive learning has been widely explored in VLM pre-training, which designs contrastive objectives for learning discriminative image-text features [10], [64], [113].

1) Image Contrastive Learning: This pre-training objective aims to learn discriminative features in image modality, which often serves as an auxiliary objective for fully exploiting the image data potential. For example, SLIP [64] employs a standard infoNCE loss defined in (1) for learning discriminative image features.

2) Image-Text Contrastive Learning: Image-text contrast aims to learn vision-language correlation by contrasting image-text pairs, i.e., pulling the embeddings of paired images and texts close while pushing others faraway [10]. For example, CLIP [10] employs a symmetrical image-text infoNCE loss in (2) which measures the image-text similarity by a dot-product between image and text embeddings in Fig. 6. The pre-trained VLM hence learns image-text correlation which allows zero-shot predictions in downstream visual recognition tasks.

Inspired by the great success of CLIP, many studies improve the symmetrical image-text infoNCE loss from different perspectives. For example, ALIGN [17] scales up the VLM pre-training with large-scale (i.e., 1.8 billions) but noisy image-text pairs with noise-robust contrastive learning. Several studies [112], [113], [114] instead explore data-efficient VLM pre-training with much less image-text pairs. For example, DeCLIP [113] introduces nearest-neighbor supervision to utilize the information from similar pairs, enabling

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

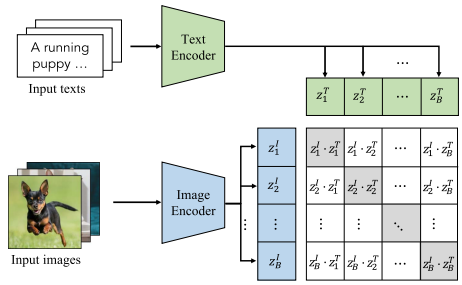

Fig. 6. Illustration of the image-text contrastive learning in CLIP [10]. Figure is reproduced from [10].

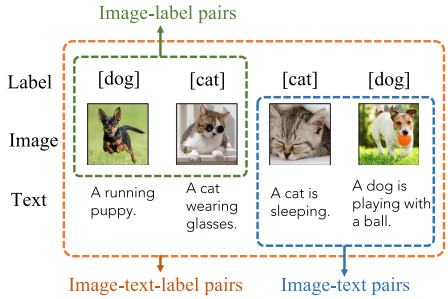

Fig. 7. Illustration of the image-text-label space proposed in UniCL [65]. Figure is reproduced from [65].

effective pre-training on limited data. OTTER [112] employs optimal transport to pseudo-pair images and texts reducing the required training data greatly. ZeroVL [114] exploits limited data resource via debiased data sampling and data augmentation with coin flipping mixup.

Another line of follow-up studies [18], [116], [129] aim for comprehensive vision-language correlation modelling by performing image-text contrastive learning across various semantic levels. For example, FILIP [18] introduces region-word alignment into contrastive learning, enabling to learn fine-grained vision-language corresponding knowledge. Pyramid-CLIP [116] constructs multiple semantic levels and performs both cross-level and peer-level contrastive learning for effective VLM pre-training.

Besides, several recent studies further improve by augmenting image-text pairs [125], [126], [127], [128]. For example, LA-CLIP [126] and ALIP [127] employ large language models to augment synthetic captions for given images while RA-CLIP [125] retrieves relevant image-text pairs for image-text pair augmentation. To facilitate efficient communications across data modalities, [44] and [43] attempt to unify vision and language learning in a single encoder.

3) Image-Text-Label Contrastive Learning: This type of pre-training introduces image classification labels [65] into the image-text contrast as defined in (4), which encodes image, text and classification labels into a shared space as shown in Fig. 7. It exploits both supervised pre-training with image labels and unsupervised VLM pre-training with image-text pairs. As reported in UniCL [65], such pre-training allows learning both discriminative and task-specific (i.e., image classification) features simultaneously. The ensuing work in [115] scales UniCL

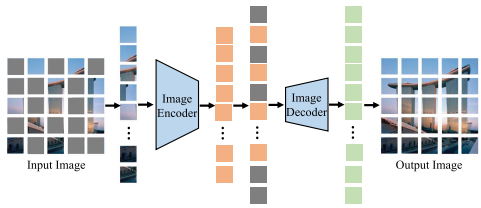

Fig. 8. Illustration of masked image modelling [66]. Figure is reproduced from [66].

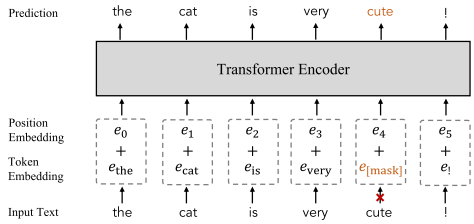

Fig. 9. Illustration of masked language modelling [14].

with around 900 M image-text pairs, leading to outstanding performance in various downstream recognition tasks.

4) Discussion: Contrastive objectives enforce positive pairs to have similar embeddings in contrast to negative pairs. They encourage VLMs to learn discriminative vision and language features [10], [17], where more discriminative features generally lead to more confident and accurate zero-shot predictions. However, the contrastive objective has two limitations: (1) Joint optimizing positive and negative pairs is complicated and challenging [10], [17]; (2) it involves a heuristic temperature hyper-parameter for controlling the feature discriminability as described in Section III-B1.

B. VLM Pre-Training With Generative Objectives

Generative VLM pre-training learns semantic knowledge by learning to generate images or texts via masked image modelling, masked language modelling, masked cross-modal modelling and image-to-text generation.

1) Masked Image Modelling: This pre-training objective guides to learn image context information by masking and reconstructing images as defined in (6). In Masked Image Modelling (e.g., MAE [41] and BeiT [70]), certain patches in an image are masked and the encoder is trained to reconstruct them conditioned on unmasked patches as shown in Fig. 8. For example, FLAVA [42] adopts rectangular block masking as in BeiT [70], while KELIP [120] and SegCLIP [46] follow MAE to mask out a large portion of patches (i.e., 75 %) in training.

2) Masked Language Modelling: Masked language modelling, a widely-adopted pre-training objective in NLP as defined in (7), also demonstrates its effectiveness in text feature learning in VLM pre-training. It works by masking a fraction of tokens in each input text and training networks to predict the masked tokens as illustrated in Fig. 9. Following [14], FLAVA [42] masks out 15% text tokens and reconstructs them from the rest tokens for modelling cross-word correlation. FIBER [71] adopts

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

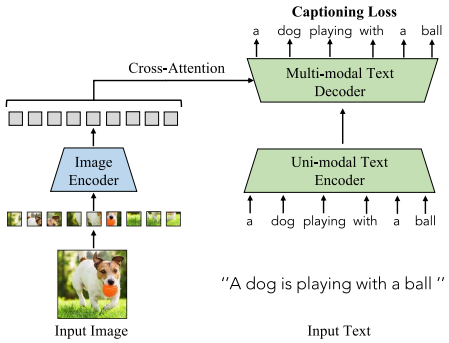

Fig. 10. Simplified illustration of image-to-caption generation in COCA [19]. Figure is reproduced based on [19].

masked language modelling [14] as one of the VLM pre-training objectives to extract better language features.

3) Masked Cross-Modal Modelling: Masked cross-modal modelling masks and reconstructs both image patches and text tokens jointly as defined in (8), which inherits the benefits of both masked image modelling and masked language modelling. It works by masking a certain percentage of image patches and text tokens and training VLMs to reconstruct them based on the embeddings of unmasked image patches and text tokens. For example, FLAVA [42] masks ∼40% image patches as in [70] and 15% text tokens as in [14], and then employs a MLP to predict masked patches and tokens, capturing rich vision-language correspondence information.

4) Image-to-Text Generation: Image-to-text generation aims to generate descriptive texts for a given image for capturing fine-grained vision-language correlation by training VLMs to predict tokenized texts. It first encodes an input image into intermediate embeddings and then decodes them into descriptive texts as defined in (9). For instance, COCA [19], NLIP [123] and PaLI [83] train VLMs with the standard encoder-decoder architecture and image captioning objectives as shown in Fig. 10.

5) Discussion: Generative objectives work by cross-modal generation or masked image/language/cross-modal modelling, encouraging VLMs to learn rich vision, language and vision-language contexts for better zero-shot predictions. Hence, generative objectives are generally adopted as additional objectives above other VLM pre-training objectives for learning rich context information [19], [42], [113].

C. VLM Pre-Training With Alignment Objectives

Alignment objectives enforce VLMs to align paired images and texts by learning to predict whether the given text describes the given image correctly. It can be broadly categorized into global image-text matching and local region-word matching for VLM pre-training.

1) Image-Text Matching: Image-text matching models global image-text correlation by directly aligning paired images and texts as defined in (10). For example, given a batch of image-text pairs, FLAVA [42] matches the given image with its paired text via a classifier and a binary classification loss. FIBER [71] follows [72] to mine hard negatives with pair-wise similarities for better alignment between images and texts.

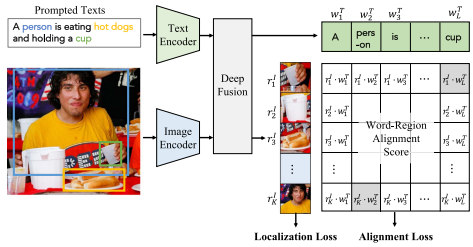

Fig. 11. Illustration of GLIP [67] that uses word-region alignment for detection. Figure is reproduced from [67].

2) Region-Word Matching: Region-word matching objective models local fine-grained vision-language correlation by aligning paired image regions and word tokens, greatly benefiting zero-shot dense predictions in object detection and semantic segmentation. For example, GLIP [67], FIBER [71] and DetCLIP [45] replace object classification logits by region-word alignment scores, i.e., the dot-product similarity between regional visual features and token-wise features as illustrated in Fig. 11.

3) Discussion: Alignment objectives learn to predict weather the given image and text data are matched or not, which are simple and easy-to-optimize and can be easily extended to model fine-grained vision-language correlation by matching image and text data locally. On the other hand, they often learn little correlation information within vision or language modality. Therefore, alignment objectives are often adopted as auxiliary losses to other VLM pre-training objectives for enhancing modelling the correlation across vision and language modalities [42], [121].

D. Summary and Discussion

In summary, VLM pre-training models the vision-language correlation with different cross-modal objectives such as image-text contrastive learning, masked cross-modal modelling, image-to-text generation and image-text/region-word matching. Various single-modal objectives have also been explored for fully exploiting the data potential of its own modality, such as masked image modelling for image modality and masked language modelling for text modality. At the other end, recent VLM pre-training focuses on learning global vision-language correlation with benefits in image-level recognition tasks such as image classification. Meanwhile, several studies [45], [46], [67], [71], [129], [130], [131] model local fine-grained vision-language correlation via region-word matching, aiming for better dense predictions in object detection and semantic segmentation.

VI. VLM Transfer Learning

Beyond *zero-shot prediction* that directly applies pre-trained VLMs on downstream tasks without fine-tuning, transfer learning has been studied recently which adapts VLMs to fit downstream tasks via prompt tuning [31], [132], feature adapter [33], [34], etc. This section presents the motivation of transfer learning for pre-trained VLMs, the common transfer-learning setup, and three transfer learning approaches including prompt tuning methods, feature adapter methods and other methods.

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
SUMMARY OF VLM TRANSFER LEARNING METHODS

Method	Category	Setup	Contribution
CoOp [31] [code]	TPT	Few-shot Sup.	Introduce context optimization with learnable text prompts for VLM transfer learning.
CoCoOp [32] [code]	TPT	Few-shot Sup.	Propose conditional text prompting to mitigate overfitting in VLM transfer learning.
SubPT [132] [code]	TPT	Few-shot Sup.	Propose subspace text prompt tuning to mitigate overfitting in VLM transfer learning.
LASP [133]	TPT	Few-shot Sup.	Propose to regularize the learnable text prompts with the hand-engineered prompts.
ProDA [134]	TPT	Few-shot Sup.	Propose prompt distribution learning that captures the distribution of diverse text prompts.
VPT [135]	TPT	Few-shot Sup.	Propose to model the text prompt learning with instance-specific distribution.
ProGrad [136] [code]	TPT	Few-shot Sup.	Present a prompt-aligned gradient technique for preventing knowledge forgetting.
CPL [137] [code]	TPT	Few-shot Sup.	Employ counterfactual generation and contrastive learning for text prompt tuning.
PLOT [138] [code]	TPT	Few-shot Sup.	Introduce optimal transport to learn multiple comprehensive text prompts.
DualCoOp [139] [code]	TPT	Few-shot Sup.	Introduce positive and negative text prompt learning for multi-label classification.
TaI-DPT [140] [code]	TPT	Few-shot Sup.	Introduce a double-grained prompt tuning technique for multi-label classification
SoftCPT [141] [code]	TPT	Few-shot Sup.	Propose to fine-tune VLMs on multiple downstream tasks simultaneously.
DenseClip [142] [code]	TPT	Supervised	Propose a language-guided fine-tuning technique for dense visual recognition tasks.
UPL [143] [code]	TPT	Unsupervised	Propose unsupervised prompt learning with self-training for VLM transfer learning.
TPT [144] [code]	TPT	Unsupervised	Propose test-time prompt tuning that learns adaptive prompts on the fly.
KgCoOp [145] [code]	TPT	Few-shot Sup.	Introduce knowledge-guided prompt tuning to improve the generalization ability.
ProTeCt [146]	TPT	Few-shot Sup.	Propose a prompt tuning technique to improve consistency of model predictions.
VP [147] [code]	VPT	Supervised	Investigate the efficacy of visual prompt tuning for VLM transfer learning.
RePrompt [148]	VPT	Few-shot Sup.	Introduce retrieval mechanisms to leverage knowledge from downstream tasks.
UPT [149] [code]	TPT, VPT	Few-shot Sup.	Propose a unified prompt tuning that jointly optimizes text and image prompts.
MVLPT [150][code]	TPT, VPT	Few-shot Sup.	Incorporate multi-task knowledge into text and image prompt tuning.
MaPLE [151][code]	TPT, VPT	Few-shot Sup.	Propose multi-modal prompt tuning with a mutual promotion strategy.
CAVPT [152][code]	TPT, VPT	Few-shot Sup.	Introduce class-aware visual prompt for concentrating more on visual concepts.
Clip-Adapter [33][code]	FA	Few-shot Sup.	Introduce an adapter with residual feature blending for efficient VLM transfer learning.
Tip-Adapter [34][code]	FA	Few-shot Sup.	Propose to build a training-free adapter with the embeddings of few labelled images.
SVL-Adpter [153][code]	FA	Few-shot Sup.	Introduce a self-supervised adapter by performing self-supervised learning on images.
SuS-X [154][code]	FA	Unsupervised	Propose a training-free name-only transfer learning paradigm with curated support sets.
CLIPPR [155][code]	FA	Unsupervised	Leverage the label distribution priors for adapting pre-trained VLMs.
SgVA-CLIP [156]	TPT, FA	Few-shot Sup.	Propose a semantic-guided visual adapter to generate discriminative adapted features.
VT-Clip [157]	CA	Few-shot Sup.	Introduce visual-guided attention that semantically aligns text and image features.
CALIP [158] [code]	CA	Unsupervised	Propose parameter-free attention for the communication between visual and textual features.
TaskRes [159] [code]	CA	Few-shot Sup.	Propose a technique for better learning old VLM knowledge and new task knowledge.
CuPL [160]	LLM	Unsupervised	Employ large language models to generate customized prompts for VLMs.
VCD [161]	LLM	Unsupervised	Employ large language models to generate captions for VLMs.
Wise-FT [162][code]	FT	Supervised	Propose ensemble-based fine-tuning by combining the fine-tuned and original VLMs.
MaskClip [163][code]	AM	Unsupervised	Propose to extract dense features by modifying the image encoder architecture.
MUST [164][code]	Self-training	Unsupervised	Propose masked unsupervised self-training for unsupervised VLM transfer learning.

TPT: text-prompt tuning; VPT: visual-prompt tuning; FA: feature adapter; CA: cross-attention; FT: fine-tuning; AM: architecture modification; LLM: large-language model. [code] directs to code websites.

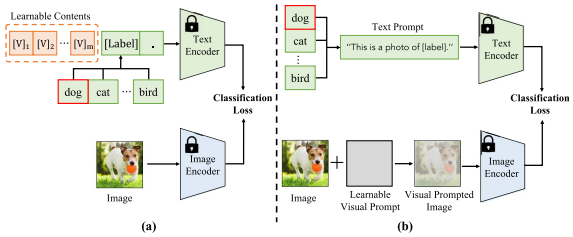

Fig. 12. Illustration of text prompt learning [31] in (a) and visual prompt learning [147] in (b).

A. Motivation of Transfer Learning

Although pre-trained VLMs have demonstrated strong generalization capability, they often face two types of gaps while applied to various downstream tasks: 1) the gaps in image and text distributions, e.g., an downstream dataset may have task-specific image styles and text formats; 2) the gaps in training objectives, e.g., VLMs are generally trained with task-agnostic objectives and learn general concepts while downstream tasks often involve task-specific objectives such as coarse or fine-grained classification, region or pixel-level recognition, etc.

B. Common Setup of Transfer Learning

Three transfer setups have been explored for mitigating the domain gaps described in Section VI-A, including supervised transfer, few-shot supervised transfer and unsupervised transfer. Supervised transfer employs all labelled downstream data for fine-tuning the pre-trained VLMs, while few-shot supervised transfer is more annotation efficient which just uses a small amount of labelled downstream samples. Differently, unsupervised transfer uses unlabelled downstream data for fine-tuning VLMs. It is thus more challenging but more promising and efficient for VLM transfer.

C. Common Transfer Learning Methods

As shown in Table IV, we broadly group existing VLM transfer methods into three categories including prompt tuning approaches, feature adapter approaches, and others.

1) Transfer Via Prompt Tuning: Inspired by the "prompt learning" in NLP [165], many VLM prompt learning methods have been proposed for adapting VLMs to fit downstream tasks by finding optimal prompts without fine-tuning the entire VLM. Most existing studies follow three approaches by text prompt tuning, visual prompt tuning, and text-visual prompt tuning.

Transfer with Text Prompt Tuning: Different from prompt engineering [165] that manually designs text prompts for each task, text prompt tuning explores more effective and efficient learnable text prompts with several labelled downstream samples for each class. For example, CoOp [31] explores context optimization to learn context words for a single class name with learnable word vectors. It expands a category word `[label]` into a sentence '$[V]_1$, $[V]_2$,..., $[V]_m$ `[label]`', where [V] denotes the learnable word vectors that are optimized by minimizing the classification loss with the downstream samples as shown in Fig. 12(a). To mitigate the overfitting due to limited downstream samples in prompt learning, CoCoOp [32] explores conditional context optimization that generates a specific prompt for each image. SubPT [132] designs subsapce prompt tuning to improve the generalization of learned prompts. LASP [133] regularizes

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

learnable prompts with hand-engineered prompts. VPT [135] models text prompts with instance-specific distribution with better generalization on downstream tasks. KgCoOp [145] enhances the generalization of unseen class by mitigating the forgetting of textual knowledge.

In addition, SoftCPT [141] fine-tunes VLMs on multiple few-shot tasks simultaneously for benefiting from multi-task learning. PLOT [138] employs optimal transport to learn multiple prompts to describe the diverse characteristics of a category. DualCoOp [139] and TaI-DP [140] transfer VLMs to multi-label classification tasks, where DualCoOp adopts both positive and negative prompts for multi-label classification while TaI-DP introduces double-grained prompt tuning for capturing both coarse-grained and fine-grained embeddings. DenseCLIP [142] explores language-guided fine-tuning that employs visual features to tune text prompts for dense prediction [55], [56]. ProTeCt [146] improves the consistency of model predictions for hierarchical classification task.

Beyond supervised and few-shot supervised prompt learning, recent studies explore unsupervised prompt tuning for better annotation efficiency and scalability. For instance, UPL [143] optimizes learnable prompts with self-training on selected pseudo-labeled samples. TPT [144] explores test-time prompt tuning to learn adaptive prompts from a single downstream sample.

Transfer with Visual Prompt Tuning: Unlike text prompt tuning, visual prompt tuning [148], [166] transfers VLMs by modulating the input of image encoder as shown in Fig. 12(b). For example, VP [147] adopts learnable image perturbations v to modify the input image x^I by $x^I + v$, aiming to adjust v to minimize a recognition loss. RePrompt [148] integrates retrieval mechanisms into visual prompt tuning, allowing leveraging the knowledge from downstream tasks. Visual prompt tuning enables pixel-level adaptation to downstream tasks, benefiting them greatly especially for dense prediction tasks.

Transfer with Text-Visual Prompt Tuning aims to modulate the text and image inputs simultaneously, benefiting from joint prompt optimization on multiple modalities. For example, UPT [149] unifies prompt tuning to jointly optimize text and image prompts, demonstrating the complementary nature of the two prompt tuning tasks. MVLPT [150] explores multi-task vision-language prompt tuning to incorporate cross-task knowledge into text and image prompt tuning. MAPLE [151] conducts multi-modal prompt tuning by aligning visual prompts with their corresponding language prompts, enabling a mutual promotion between text prompts and image prompts. CAVPT [152] introduces a cross attention between class-aware visual prompts and text prompts, encouraging the visual prompts to concentrate more on visual concepts.

Discussion: Prompt tuning enables parameter-efficient VLM transfer by modifying input texts/images with a few learnable text/image prompts. It is simple and easy-to-implement, and requires little extra network layers or complex network modifications. Therefore, prompt tuning allows adapting VLMs in a black-box manner, which has clear advantages in transferring VLMs that involve concerns in intellectual property. However, it still suffers from several limitations such as the low flexibility by following the manifold of the original VLMs in prompting [31].

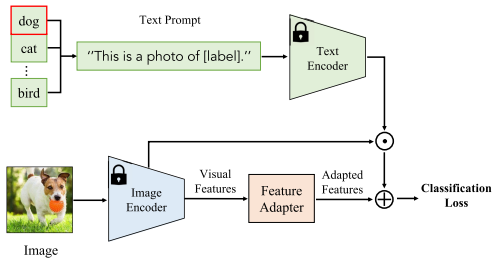

Fig. 13. Illustration of feature adapter [33].

2) Transfer Via Feature Adaptation: Feature adaptation fine-tunes VLMs to adapt image or text features with an additional light-weight feature adapter [167]. For example, Clip-Adapter [33] inserts several trainable linear layers after CLIP's language and image encoders and optimizes them while keeping CLIP architecture and parameters frozen as illustrated in Fig. 13. Tip-Adapter [34] presents a training-free adapter that directly employs the embeddings of few-shot labelled images as the adapter weights. SVL-Adapter [153] designs a self-supervised adapter which employs an additional encoder for self-supervised learning on input images. In summary, feature adapter adapts image and text features to fit VLMs to downstream data, which provides a promising alternative to prompt tuning for VLMs transfer.

Discussion: Feature adaptation adapts VLMs by modifying image and text features with an additional light-weight feature adapter. It is flexible and effective as its architecture and the insertion manner allow tailoring flexibly for different downstream tasks. Therefore, feature adaptation has clear advantages in adapting VLMs to work on very different and complex downstream tasks [168], [169], [170], [171]. On the other hand, it requires modifying network architecture and thus can not handle VLMs that have concerns in intellectual property.

3) Other Transfer Methods: Several studies transfer VLMs by direct fine-tuning [162], architecture modification [163], and cross attention [157], [158]. Specifically, Wise-FT [162] combines the weights of a fine-tuned VLM and the original VLM for learning new information from downstream tasks. MaskCLIP [163] extracts dense image features by modifying the architecture of the CLIP image encoder. VT-CLIP [157] introduces visual-guided attention to semantically correlate text features with downstream images, leading to a better transfer performance. CALIP [158] introduces parameter-free attention for effective interaction and communication between visual and text features, leading to text-aware image features and visual-guided text features. TaskRes [159] directly tunes text-based classifier to exploit the old knowledge in the pre-trained VLM. CuPL [160] and VCD [161] employ large language models, e.g., GPT_3 [172], to augment text prompts for learning rich discriminative text information.

D. Summary and Discussion

In summary, prompt tuning and feature adapter are two major approaches for VLM transfer which work by modifying the input text/image and adapting image/text features, respectively. In addition, both approaches introduce very limited parameters

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

while freezing the original VLMs, leading to efficient transfer. Further, while most studies follow few-shot supervised transfer [31], [32], [132], [134], recent studies show that unsupervised VLM transfer can achieve competitive performance on various tasks [143], [144], [160], inspiring more research on unsupervised VLM transfer.

VII. VLM Knowledge Distillation

As VLMs capture generalizable knowledge that covers a wide range of visual and text concepts, several studies explore how to distil the general and robust VLM knowledge while tackling complex dense prediction tasks such as object detection and semantic segmentation. This section presents the motivation of distilling knowledge from VLMs as well as two groups of knowledge distillation studies on the tasks of semantic segmentation and object detection.

A. Motivation of Distilling Knowledge From VLMs

Different from VLM transfer that generally keeps the original VLM architecture intact in transfer [31], [132], [136], VLM knowledge distillation distils general and robust VLM knowledge to task-specific models without the restriction of VLM architecture, benefiting task-specific designs while tackling various dense prediction tasks [36], [173], [174]. For example, knowledge distillation allows transferring the general VLM knowledge to tackle detection tasks while taking the advantages of state-of-the-art detection architectures such as Faster R-CNN [55] and DETR [62].

B. Common Knowledge Distillation Methods

As VLMs are generally pre-trained with architectures and objectives designed for image-level representation, most VLM knowledge distillation methods focus on transferring image-level knowledge to region- or pixel-level tasks such as object detection and semantic segmentation. Table I in Appendix D, available online shows a list of VLM knowledge distillation methods.

1) Knowledge Distillation for Object Detection: Open-vocabulary object detection [175] aims to detect objects described by arbitrary texts, i.e., objects of any categories beyond the base classes. As VLMs like CLIP are trained with billion-scale image-text pairs that cover very broad vocabulary, many studies explore to distill VLM knowledge to enlarge the detector vocabulary. For example, ViLD [36] distills VLM knowledge to a two-stage detector whose embedding space is enforced to be consistent with that of CLIP image encoder. Following ViLD, HierKD [176] explores hierarchical global-local knowledge distillation, and RKD [177] explores region-based knowledge distillation for better aligning region-level and image-level embeddings. ZSD-YOLO [178] introduces self-labelling data augmentation for exploiting CLIP for better object detection. OADP [179] preserves proposal features while transferring contextual knowledge. BARON [180] uses neighborhood sampling to distill a bag of regions instead of individual regions. RO-ViT [181] distills regional information from VLMs for open-vocabulary detection.

Another line of research explores VLM distillation via prompt learning [165]. For example, DetPro [37] introduces a detection prompt technique for learning continuous prompt representations for open-vocabulary object detection. PromptDet [182] introduces regional prompt learning for aligning word embeddings with regional image embeddings. Additionally, several studies [183], [184], [185], [186], [187] explore VLM-predicted pseudo labels to improve object detectors. For example, PB-OVD [183] trains object detectors with VLM-predicted pseudo bounding boxes while XPM [184] introduces a robust cross-modal pseudo-labeling strategy that employs VLM-generated pseudo masks for open-vocabulary instance segmentation. P ^{3} OVD [185] exploits prompt-driven self-training that refines the VLM-generated pseudo labels with fine-grained prompt tuning.

2) Knowledge Distillation for Semantic Segmentation: Knowledge distillation for open-vocabulary semantic segmentation leverages VLMs to enlarge the vocabulary of segmentation models, aim to segment pixels described by arbitrary texts (i.e., any categories of pixels beyond base classes). For example, [35], [186], [187] achieve open-vocabulary semantic segmentation by first class-agnostic segmentation by grouping pixels into multiple segments and then segment recognition with CLIP. CLIPSeg [188] introduces a lightweight transformer decoder to extend CLIP for semantic segmentation. LSeg [189] maximizes the correlation between CLIP text embeddings and pixel-wise image embedding encoded by segmentation models. ZegCLIP [174] employs CLIP to generate semantic masks and introduces a relationship descriptor to mitigate overfitting on base classes. MaskCLIP+ [163] and SSIW [190] distill knowledge with VLM-predicted pixel-level pseudo labels. FreeSeg [191] generates mask proposals first and then performs zero-shot classification for them.

Knowledge distillation for weakly-supervised semantic segmentation aims to leverage both VLMs and weak supervision (e.g., image-level labels) for semantic segmentation. For example, CLIP-ES [192] employs CLIP to refine the class activation map by deigning a softmax function and a class-aware attention-based affinity module for mitigating the category confusion issue. CLIMS [193] employs CLIP knowledge to generate high-quality class activation maps for better weakly-supervised semantic segmentation.

C. Summary and Discussion

In summary, most VLM studies explore knowledge distillation over two dense visual recognition tasks, namely, object detection and semantic segmenting, where those for the former aim to better align image-level and object-level representations while those for the latter focus on tackling the mismatch between image-level and pixel-level representations. They can also be categorized based on their methodology, including feature-space distillation that enforces embedding consistency between VLM's encoder and the detection (or segmentation) encoder and pseudo-labelling distillation that employs VLM-generated pseudo labels to regularize detection or segmentation models. Moreover, compared with VLM transfer, VLM knowledge distillation has clearly better flexibility of allowing different downstream networks regardless of the original VLMs.

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

TABLE V
PERFORMANCE OF VLM PRE-TRAINING METHODS OVER ZERO-SHOT PREDICTION SETUP ON IMAGE CLASSIFICATION TASKS

Methods	Image encoder	Text encoder	Data Size	ImageNet-1k [40]	CIFAR-10 [23]	CIFAR-100 [23]	Food101 [22]	sun397 [24]	Cars [25]	Aircraft [96]	DTD [99]	Pets [26]	caltech101 [89]	flowers102 [91]
CLIP [10]	ViT-L/14	Transformer	400M	76.2	95.7	77.5	93.8	68.4	78.8	37.2	55.7	93.5	92.8	78.3
ALIGN [17]	EfficientNet	BERT	1.8B	76.4	-	-	-	-	-	-	-	-	-	-
OTTER [112]	FBNetV3-C	DeCLUTR-Sci	3M	-	-	-	-	-	-	-	-	-	-	-
DeCLIP [113]	REGNET-Y	BERT	88M	73.7	-	-	-	-	-	-	-	-	-	-
ZeroVL [114]	ViT-B/16	BERT	100M	-	-	-	-	-	-	-	-	-	-	-
FILIP [18]	ViT-L/14	Transformer	340M	77.1	95.7	75.3	92.2	73.1	70.8	60.2	60.7	92.0	93.0	90.1
UniCL [65]	Swin-tiny	Transformer	16.3M	71.3	-	-	-	-	-	-	-	-	-	-
Florence [115]	CoSwin	RoBERT	900M	83.7	94.6	77.6	95.1	77.0	93.2	55.5	66.4	95.9	94.7	86.2
SLIP [64]	ViT-L	Transformer	15M	47.9	87.5	54.2	69.2	56.0	9.0	9.5	29.9	41.6	80.9	60.2
PyramidCLIP [116]	ResNet50	T5	143M	47.8	81.5	53.7	67.8	65.8	65.0	12.6	47.2	83.7	81.7	65.8
Chinese CLIP [117]	ViT-L/14	CNRoberta	200M	-	96.0	79.7	-	-	-	26.2	51.2	-	-	-
LiT [118]	ViT-g/14	-	4B	85.2	-	-	-	-	-	-	-	-	-	-
AltCLIP [119]	ViT-L/14	Transformer	2M	74.5	-	-	-	-	-	-	-	-	-	-
FLAVA [42]	ViT-B/16	ViT-B/16	70M	-	-	-	-	-	-	-	-	-	-	-
KELIP [120]	ViT-B/32	Transformer	1.1B	62.6	91.5	68.6	79.5	-	75.4	-	51.2	-	-	-
COCA [19]	ViT-G/14	-	4.8B	86.3	-	-	-	-	-	-	-	-	-	-
nCLIP [121]	ViTB/16	Transformer	35M	48.8	83.4	54.5	65.8	59.9	18.0	5.8	57.1	33.2	73.9	50.0
K-lite [122]	CoSwin	RoBERT5	813M	85.8	-	-	-	-	-	-	-	-	-	-
NLIP [123]	ViT-B/16	BART	26M	47.4	81.9	47.5	59.2	58.7	7.8	7.5	32.9	39.2	79.5	54.0
UniCLIP [84]	ViT-B/32	Transformer	30M	54.2	87.8	56.5	64.6	61.1	19.5	4.7	36.6	69.2	84.0	8.0
PaLI [83]	ViT-e	mT5	12B	85.4	-	-	-	-	-	-	-	-	-	-
CLIPPO [43]	ViT-L/16	ViT-L/16	12B	70.5	-	-	-	-	-	-	-	-	-	-
OneR [44]	ViT-L/16	ViT-L/16	4M	27.3	-	31.4	-	-	-	-	-	-	-	-
RA-CLIP [125]	ViT-B/32	BERT	15M	53.5	89.4	62.3	43.8	46.5	-	-	25.6	-	76.9	-
LA-CLIP [126]	ViT-B/32	Transformer	400M	64.4	92.4	73.0	79.7	64.9	81.9	20.8	55.4	87.2	91.8	70.3
ALIP [127]	ViT-B/32	Transformer	15M	40.3	83.8	51.9	45.4	47.8	3.4	2.7	23.2	30.7	74.1	54.8
GrowCLIP [128]	ViT-B/16	Transformer	12M	36.1	60.7	28.3	42.5	45.5	-	-	17.3	-	71.9	23.3

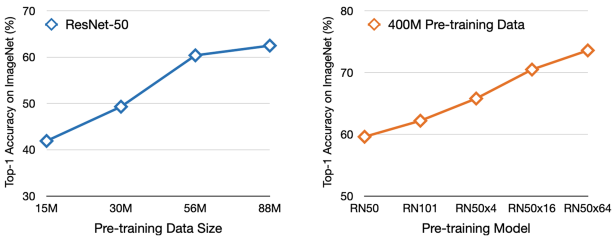

Fig. 14. Performance versus data size and model size. It shows that scaling up either the pre-training data [113] or the pre-training model [10] benefits VLM consistently.

VIII. PERFORMANCE COMPARISON

In this section, we compare, analyze and discuss the VLM pre-training, VLM transfer learning, and VLM knowledge distillation methods as reviewed in Sections V-VII.

A. Performance of VLM Pre-Training

As discussed in Section III-D, *zero-shot prediction* as one widely-adopted evaluation setup assesses VLM generalization over unseen tasks without task-specific fine-tuning. This subsection presents the performance of *zero-shot prediction* over different visual recognition tasks including image classification, object detection, and semantic segmentation.

Table V shows evaluations on 11 widely adopted image classification tasks. Note it shows the best VLM performance as VLM pre-training often have different implementations. Three conclusions can be drawn from Table V as well as Fig. 14: 1) VLM performance is usually up to the size of training data. As shown in the first graph in Fig. 14, scaling up the pre-training data leads to consistent improvements; 2) VLM performance is usually up to the model size. As shown in the second graph, with the same pre-training data, scaling up model sizes improves the VLM performance consistently; 3) With large-scale image-text training data, VLMs can achieve superior zero-shot performance on various downstream tasks. As Table V shows, COCA [19] achieves state-of-the-art performance on ImageNet, and FILIP [18] performs well consistently across 11 tasks.

The superior generalization of VLMs is largely attributed to three factors: 1) Big data - as image-text pairs are almost infinitely available on the Internet, VLMs are usually trained with millions or billions of image and text samples that cover very broad visual and language concepts, leading to strong generalization capability; 2) Big model - compared with traditional visual recognition models, VLMs generally adopt much larger models (e.g., ViT-G in COCA [19] with 2B parameters) that provide great capacity for effective learning from Big Data; 3) Task-agnostic learning - the supervision in VLM pre-training is usually general and task-agnostic. Compared with task-specific labels in traditional visual recognition, the texts in image-text pairs provide task-agnostic, diverse and informative language supervision which help train generalizable models that works well across various downstream tasks.

Note several studies [45], [46], [67], [71], [129], [131] investigate VLM pre-training for object detection and semantic segmentation with local VLM pre-training objectives such as region-word matching [67]. Tables VI and VII summarize *zero-shot prediction* performance on object detection and semantic segmentation tasks. We can observe that VLMs enable effective zero-shot prediction on both dense prediction tasks. Note the results in Tables VI and VII may not be aligned with the conclusions in previous paragraphs, largely because this field of research is under-explored with very limited VLMs on dense visual tasks.

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

TABLE VI
PERFORMANCE OF VLM PRE-TRAINING METHODS OVER ZERO-SHOT PREDICTION SETUP ON SEGMENTATION TASKS

Method	Image encoder	Text encoder	Data size	VOC [90]	PASCAL C. [109]	COCO [106]
GroupVit [129]	ViT	Transformer	26M	52.3	22.4	-
SegClip [46]	ViT	Transformer	3.4M	52.6	24.7	26.5

TABLE VII
PERFORMANCE OF VLM PRE-TRAINING METHODS OVER ZERO-SHOT PREDICTION SETUP ON DETECTION TASKS

Method	Image encoder	Text encoder	Data size	COCO [106]	LVIS [107]	LVIS Mini. [107]
RegionClip [131]	ResNet50x4	Transformer	118k	29.6	11.3	-
GLIP [67]	Swin-L	BERT	27.43M	49.8	26.9	34.3
FIBER [71]	Swin-B	RoBERTa	4M	49.3	-	32.2
DetCLIP [45]	Swin-L	BERT	2.43M	-	35.9	-

Limitations of VLMs: As discussed above, although VLMs benefit clearly while data/model size scales up, they still suffer from several limitations: (1) When data/model size keeps increasing, the performance saturates and further scaling up won't improve performance [113], [194]; (2) Adopting large-scale data in VLM pre-training necessitates extensive computation resources, e.g., 256 V100 GPUs, 288 training hours in CLIP ViT-L [10]; (3) Adopting large models introduces excessive computation and memory overheads in both training and inference.

B. Performance of VLM Transfer Learning

This section summarizes the performance of VLM transfer under the setups of supervised transfer, few-shot supervised transfer and unsupervised transfer. Table VIII shows the results on 11 widely adopted image classification datasets (e.g., EuroSAT [104], UCF101 [29]) with different backbones such as CNN backbone ResNet-50 and Transformer backbones ViT-B and ViT-L. Note Table VIII summarizes the performance of 16-shot setup for all *few-shot supervised* methods.

Three conclusions can be drawn from Table VIII. First, VLM transfer setups helps in downstream tasks consistently. For example, supervised Wise-FT, few-shot supervised CoOp and unsupervised TPT improve accuracy by 10.9%,1.7% and 0.8%, respectively, on ImageNet. As pre-trained VLMs generally suffer from domain gaps with task-specific data, VLM transfer can mitigate the domain gaps by learning from task-specific data, being labelled or unlabelled.

Second, the performance of few-shot supervised transfer lag far behind that of supervised transfer (e.g., 87.1% in WiseFT [162] and 76.6% in CuPL [160]), largely because VLMs may overfit to few-shot labelled samples with degraded generalization. Third, unsupervised transfer can perform comparably with few-shot supervised transfer (e.g., unsupervised UPL [143] outperforms 2-shot supervised CoOp [31] by 0.4%, unsupervised TPT [144] is comparable with 16-shot CoOp [31]), largely because unsupervised transfer can access massive unlabelled downstream data with much lower overfitting risks. Nevertheless, unsupervised transfer also faces several challenges such as noisy pseudo labels. We expect more studies on this promising but changeling research direction.

C. Performance of VLM Knowledge Distillation

This section presents how VLM knowledge distillation helps in the tasks of object detection and semantic segmentation. Tables IX and X show the knowledge distillation performance on the widely used detection datasets (e.g., COCO [106] and LVIS [107]) and segmentation datasets (e.g., PASCAL VOC [90] and ADE20k [111]), respectively. We can observe that VLM knowledge distillation brings clear performance improvement on detection and segmentation tasks consistently, largely because it introduces general and robust VLM knowledge while benefiting from task-specific designs in detection and segmentation models.

D. Summary

Several conclusions can be drawn from Tables V–X. Regarding *performance*, VLM pre-training achieves remarkable zero-shot prediction on a wide range of image classification tasks due to its well-designed pre-training objectives. Nevertheless, the development of VLM pre-training for dense visual recognition tasks (on region or pixel-level detection and segmentation) lag far behind. In addition, VLM transfer has made remarkable progress across multiple image classification datasets and vision backbones. However, supervised or few-shot supervised transfer still requires labelled images, whereas the more promising but challenging unsupervised VLM transfer has been largely neglected.

Regarding *benchmark*, most VLM transfer studies adopt the same pre-trained VLM as the baseline model and perform evaluations on the same downstream tasks, which facilitates benchmarking greatly. They also release their codes and do not require intensive computation resources, easing reproduction and benchmarking greatly. Differently, VLM pre-training has been studied with different data (e.g., CLIP [10], LAION400M [21] and CC12M [79]) and networks (e.g., ResNet [6], ViT [57], Transformer [58] and BERT [14]), making fair benchmarking a very challenging task. Several VLM pre-training studies also use non-public training data [10], [18], [83] or require intensive computation resources (e.g., 256 V100 GPUs in [10]). For VLM knowledge distillation, many studies adopt different task-specific backbones (e.g., ViLD adopts Faster R-CNN, OV-DETR uses DETR) which complicates benchmarking greatly. Hence, VLM pre-training and VLM knowledge distillation are short of certain norms in term of training data, networks and downstream tasks.

IX. FUTURE DIRECTIONS

VLM enables effective usage of web data, zero-shot prediction without any task-specific fine-tuning, and open-vocabulary visual recognition of images of arbitrary categories. It has been achieving great success with incredible visual recognition performance. In this section, we humbly share several research challenges and potential research directions that could be pursued in the future VLM study on various visual recognition tasks.

For **VLM pre-training**, there are four challenges and potential research directions as listed.

1) Fine-grained vision-language correlation modelling: With local vision-language correspondence knowledge [45], [67],

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII
PERFORMANCE OF VLM TRANSFER LEARNING METHODS ON IMAGE CLASSIFICATION TASKS

Methods	Image encoder	Setup	Average	ImageNet-1k [40]	caltech101 [89]	Pets [26]	Cars [25]	Flowers102 [91]	Food101 [22]	Aircraft [96]	SUN397 [24]	DTD [99]	EuroSAT [104]	UCF101 [29]
Baseline [143]	ResNet-50	w/o Transfer	59.2	60.3	86.1	85.8	55.6	66.1	77.3	16.9	60.2	41.6	38.2	62.7
Baseline [10]	ViT-B/16	w/o Transfer	71.7	70.2	95.4	94.1	68.6	74.8	90.6	31.1	72.2	56.4	60.6	73.5
Baseline [10]	ViT-L/14	w/o Transfer	73.7	76.2	92.8	93.5	78.8	78.3	93.8	37.2	68.4	55.7	59.6	76.9
CoOp [31]	ViT-B/16	Few-shot Sup.	71.6	71.9	93.7	94.5	68.1	74.1	85.2	28.7	72.5	54.2	68.7	67.5
CoCoOp [32]	ViT-B/16	Few-shot Sup.	75.8	73.1	95.8	96.4	72.0	81.7	91.0	27.7	78.3	64.8	71.2	77.6
SubPT [132]	ResNet50	Few-shot Sup.	66.4	63.4	91.7	91.8	60.7	73.8	81.0	20.3	70.2	54.7	54.5	68.1
LASP [133]	ViT-B/16	Few-shot Sup.	76.1	73.0	95.8	95.7	72.2	81.6	90.5	31.6	77.8	62.8	74.6	76.8
ProDA [134]	ResNet50	Few-shot Sup.	-	65.3	91.3	90.0	75.5	95.5	82.4	36.6	-	70.1	84.3	-
VPT [135]	ViT-B/16	Few-shot Sup.	77.4	73.4	96.4	96.8	73.1	81.1	91.6	34.7	78.5	67.3	77.7	79.0
ProGrad [136]	ResNet-50	Few-shot Sup.	67.9	62.1	91.5	93.4	62.7	78.7	81.0	21.9	70.3	57.8	59.0	68.5
CPL [137]	ViT-B/16	Few-shot Sup.	-	76.0	96.3	97.7	77.2	81.7	93.2	-	80.6	-	-	-
PLOT [138]	ResNet-50	Few-shot Sup.	73.9	63.0	92.2	87.2	72.8	94.8	77.1	34.5	70.0	65.6	82.2	77.3
CuPL [160]	ViT-L/14	Few-shot Sup.	-	76.6	93.4	93.8	77.6	-	93.3	36.1	61.7	-	-	-
UPL [143]	ResNet-50	Unsupervised	68.4	61.1	91.4	89.5	71.0	76.6	77.9	21.7	66.4	55.1	71.0	70.2
TPT [144]	ViT-B/16	Unsupervised	64.8	69.0	94.2	87.8	66.9	69.0	84.7	24.8	65.5	47.8	42.4	60.8
VP [147]	ViT-B/32	Few-shot Sup.	-	-	-	85.0	-	70.3	78.9	-	60.6	57.1	96.4	66.1
UPT [149]	ViT-B/16	Few-shot Sup.	76.2	73.2	96.1	96.3	71.8	81.0	91.3	34.5	78.7	65.6	72.0	77.2
MaPLE [151]	ViT-B/16	Few-shot Sup.	78.6	73.5	96.0	96.6	73.5	82.6	91.4	36.5	79.7	68.2	82.4	80.8
CAVPT [152]	ViT-B/16	Few-shot Sup.	83.2	72.5	96.1	93.5	88.2	97.6	85.0	57.9	74.3	72.6	92.1	85.3
Tip-Adapter [34]	ViT-B/16	Few-shot Sup.	-	70.8	-	-	-	-	-	-	-	-	-	-
SuS-X [154]	ResNet-50	Unsupervised	-	61.8	-	-	-	-	-	-	-	-	45.6	50.6
SgVA-CLIP [156]	ViT-B/16	Few-shot Sup.	-	73.3	-	-	-	-	-	-	76.4	-	-	-
VT-Clip [157]	ResNet-50	Few-shot Sup.	-	-	-	93.1	-	-	-	-	-	65.7	-	-
CALIP [158]	ResNet-50	Unsupervised	59.4	60.6	87.7	58.6	77.4	66.4	56.3	17.7	86.2	42.4	38.9	61.7
Wise-FT [162]	ViT-L/14	Supervised	-	87.1	-	-	-	-	-	-	-	-	-	-
KgCoOp [145]	ViT-B/16	Few-shot Sup.	74.4	70.1	94.6	93.2	71.9	90.6	86.5	32.4	71.7	58.3	71.0	78.4
ProTeCt [146]	ViT-B/16	Few-shot Sup.	69.9	-	-	-	-	-	-	-	74.5	-	-	-
RePrompt [148]	ViT-B/16	Few-shot Sup.	83.2	74.6	96.5	93.7	85.0	97.1	87.4	50.3	77.5	73.7	92.9	86.4
TaskRes [159]	ResNet-50	Few-shot Sup.	75.7	65.7	93.4	87.8	76.8	96.0	77.6	36.3	70.6	67.1	84.0	77.9
VCD [161]	ViT-B/16	Unsupervised	-	68.0	-	86.9	-	-	88.5	-	-	45.5	48.6	-

TABLE IX
PERFORMANCE OF VLM KNOWLEDGE DISTILLATION ON OBJECT DETECTION

Method	Vision-Language Model	COCO [106]			LVIS [107]			
		AP_{base}	AP_{novel}	AP	AP_r	AP_c	AP_f	AP
Baseline [36]	-	28.3	26.3	27.8	19.5	19.7	17.0	18.6
ViLD [36]	CLIP ViT-B/32	59.5	27.6	51.3	16.7	26.5	34.2	27.8
DetPro [37]	CLIP ViT-B/32	-	-	34.9	20.8	27.8	32.4	28.4
HierKD [176]	CLIP ViT-B/32	53.5	27.3	-	-	-	-	-
RKD [177]	CLIP ViT-B/32	56.6	36.9	51.0	21.1	25.0	29.1	25.9
PromptDet [182]	CLIP Transformer	-	26.6	50.6	21.4	23.3	29.3	25.3
PB-OVD [183]	CLIP Transformer	46.1	30.8	42.1	-	-	-	-
CondHead [195]	CLIP ViT-B/32	60.8	29.8	49.0	18.8	28.3	33.7	28.8
VLDet [196]	CLIP Transformer	50.6	32.0	45.8	26.3	39.4	41.9	38.1
F-VLM [197]	CLIP ResNet-50	-	28.0	39.6	32.8	-	-	34.9
OV-DETR [173]	CLIP ViT-B/32	52.7	29.4	61.0	17.4	25.0	32.5	26.6
Detic [175]	CLIP Transformer	45.0	27.8	47.1	17.8	26.3	31.6	26.8
OWL-ViT [198]	CLIP ViT-B/32	-	-	28.1	18.9	-	-	22.1
VL-PLM [199]	CLIP ViT-B/32	60.2	34.4	53.5	-	-	-	22.2
P^3OVD [185]	CLIP ResNet-50	51.9	31.5	46.6	-	-	-	10.6
RO-ViT [181]	CLIP ViT-L/16	-	33.0	47.7	32.1	-	-	34.0
BARON [180]	CLIP ResNet-50	54.9	42.7	51.7	23.2	29.3	32.5	29.5
OADP [179]	CLIP ViT-B/32	53.3	30.0	47.2	21.9	28.4	32.0	28.7

CLIP Transformer is CLIP text encoder.

TABLE X
PERFORMANCE OF VLM KNOWLEDGE DISTILLATION ON SEMANTIC SEGMENTATION TASKS

Method	Vision-Language Model	A-847 [111]	PC-459 [109]	A-150 [111]	PC-59 [109]	PAS-20 [90]	C-19 [110]
Baseline [200]	-	-	-	-	24.3	18.3	-
LSeg [35]	CLIP ResNet-101	-	-	-	-	47.4	-
ZegFormer [189]	CLIP ResNet-50	-	-	16.4	-	80.7	-
OVSeg [201]	CLIP Swin-B	9.0	12.4	29.6	55.7	94.5	-
ZSSeg [187]	CLIP ResNet-101	7.0	-	20.5	47.7	-	34.5
OpenSeg [186]	CLIP Eff-B7	6.3	9.0	21.1	42.1	-	-
ReCo [202]	CLIP ResNet-101	-	-	-	-	-	24.2
FreeSeg [191]	CLIP ViT-B/16	-	-	39.8	-	86.9	-

VLMs can better recognize patches and pixels beyond images, greatly benefiting dense prediction tasks such as object detection and semantic segmentation that play an important role in various visual recognition tasks. Given the very limited VLM studies along this direction [45], [46], [67], [71], [129], [131], we expect more research in fine-grained VLM pre-training for zero-shot dense prediction tasks.

2) Unification of vision and language learning: The advent of Transformer [57], [58] makes it possible to unify image and text learning within a single Transformer by tokenizing images and texts in the same manner. Instead of employing two separate networks as in existing VLMs [10], [17], unifying vision and language learning enables efficient communications across data modalities which can benefit both training effectiveness and training efficiency. This issue has attracted some attention [43], [44] but more efforts are needed towards more sustainable VLMs.

3) Pre-training VLMs with multiple languages: Most existing VLMs are trained with a single language (i.e., English) [10], [17], which could introduce bias in term of cultures and regions [77], [79] and hinder VLM applications in other language areas. Pre-training VLMs with texts of multiple languages [119], [120] allows learning different cultural visual characteristics for the same meaning of words but different languages [20], enabling VLMs to work efficiently and effectively across different language scenarios. We expect more research on multilingual VLMs.

4) Data-efficient VLMs: Most existing work trains VLMs with large-scale training data and intensive computations, making its sustainability a big concern. Training effective VLMs with limited image-text data can mitigate this issue greatly. For example, instead of merely learning from each image-text pair, more useful information could be learned with the supervision among image-text pairs [112], [113].

5) Pre-training VLMs with LLMs: Recent studies [126], [127] retrieve rich language knowledge from LLMs to enhance VLM pre-training. Specifically, they employ LLMs to augment the texts in the raw image-text pairs, which provides richer language

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

knowledge and helps better learn vision-language correlation. We expect more exploration of LLMs in VLM pre-training in the future research.

For **VLM Transfer Learning**, there are three challenges and potential research directions as listed.

1) Unsupervised VLM transfer: Most existing VLM transfer studies work with a supervised or few-shot supervised setup that requires labelled data, and the latter tends to overfit to the few-shot samples. Unsupervised VLM transfer allows exploring massive unlabelled data with much lower risk of overfitting. More studies on unsupervised VLM transfer are expected in the ensuing VLM studies.

2) VLM transfer with visual prompt/adapter: Most existing studies on VLM transfer focus on text prompt learning [31]. Visual prompt learning or visual adapter, which is complementary to text prompting and can enable pixel-level adaptation in various dense prediction tasks, is largely neglected. More VLM transfer studies in visual domain are expected.

3) Test-time VLM transfer: Most existing studies conduct transfer by fine-tuning VLMs on each downstream task (i.e., prompt learning), leading to repetitive efforts while facing many downstream tasks. Test-time VLM transfer allows adapting prompts on the fly during inference, circumventing the repetitive training in existing VLM transfer. We can foresee more studies on test-time VLM transfer.

4) VLM transfer with LLMs: Different from prompt engineering and prompt learning, several attempts [160], [161] exploit LLMs [172] to generate text prompts that better describe downstream tasks. This approach is automatic and requires little labelled data. We expect more exploration of LLMs in VLM transfer in the future research.

VLM knowledge distillation could be further explored from two aspects. The first is knowledge distillation from multiple VLMs that could harvest their synergistic effect by coordinating knowledge distillation from multiple VLMs. The second is knowledge distillation for other visual recognition tasks such as instance segmentation, panoptic segmentation, person re-identification etc.

X. Conclusion

Vision-language models for visual recognition enables effective usage of web data and allows zero-shot predictions without task-specific fine-tuning, which is simple to implement yet has achieved great success on a wide range of recognition tasks. This survey extensively reviews vision-language models for visual recognition from several perspectives, including background, foundations, datasets, technical approaches, benchmarking, and future research directions. The comparative summary of the VLM datasets, approaches, and performance in tabular forms provides a clear big picture of the recent development in VLM pre-training which will greatly benefit the future research along this emerging but very promising research direction.

References

[1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.

[2] G. Cheng, X. Xie, J. Han, L. Guo, and G. -S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020.

[3] H. A. Pierson and M. S. Gashler, "Deep learning in robotics: A review of recent research," *Adv. Robot.*, vol. 31, no. 16, pp. 821–835, 2017.

[4] A. Krizhevsky et al., "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 770–778, 2016.

[7] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, 2009, Art. no. 1883.

[8] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.

[9] A. Mathur and G. M. Foody, "Multiclass and binary SVM classification: Implications for training and classification users," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 2, pp. 241–245, Apr. 2008.

[10] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[11] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[12] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.

[13] T. Chen et al., "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[14] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv: 1810.04805*.

[15] A. Radford et al., "Improving language understanding by generative pre-training," *OpenAI Blog*, 2018. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[16] A. Radford et al., "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, 2019, Art. no. 9.

[17] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.

[18] L. Yao et al., "Filip: Fine-grained interactive language-image pre-training," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–13.

[19] J. Yu et al., "COCA: Contrastive captioners are image-text foundation models," 2022, *arXiv:2205.01917*.

[20] C. Schuhmann et al., "LAION-5B: An open large-scale dataset for training next generation image-text models," 2022, *arXiv:2210.08402*.

[21] C. Schuhmann et al., "LAION-400M: Open dataset of clip-filtered 400 million image-text pairs," 2021, *arXiv:2111.02114*.

[22] L. Bossard et al., "Food-101–mining discriminative components with random forests," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 446–461.

[23] A. Krizhevsky and H. Geoffrey, "Learning multiple layers of features from tiny images," Univ. Toronto, Tech. Rep., 2009. [Online]. Available: https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf

[24] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3485–3492.

[25] J. Krause et al., "Collecting a large-scale dataset of fine-grained cars," Stanford AI Lab, 2013. [Online]. Available: https://ai.stanford.edu/~jkrause/papers/fgvc13.pdf

[26] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3498–3505.

[27] D. Kiela et al., "The hateful memes challenge: Detecting hate speech in multimodal memes," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 2611–2624.

[28] A. Miech et al., "RareAct: A video dataset of unusual interactions," 2020, *arXiv: 2008.01018*.

[29] K. Soomro et al., "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[30] J. Carreira et al., "A short note on the kinetics-700 human action dataset," 2019, *arXiv: 1907.06987*.

[31] K. Zhou et al., "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

[32] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16795–16804.

[33] P. Gao et al., "Clip-adapter: Better vision-language models with feature adapters," 2021, *arXiv:2110.04544*.

[34] R. Zhang et al., "Tip-adapter: Training-free clip-adapter for better vision-language modeling," 2021, *arXiv:2111.03930*.

[35] J. Ding, N. Xue, G. Xia, and D. Dai, "Decoupling zero-shot semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11583–11592.

[36] X. Gu et al., "Open-vocabulary object detection via vision and language knowledge distillation," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–14.

[37] Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, and G. Li, "Learning to prompt for open-vocabulary object detection with vision-language model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14064–14073.

[38] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.

[39] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[40] J. Deng, W. Dong, R. Socher, L. -J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[41] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15979–15988.

[42] A. Singh et al., "FLAVA: A foundational language and vision alignment model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15617–15629.

[43] M. Tschannen et al., "Image-and-language understanding from pixels only," 2022, *arXiv:2212.08045*.

[44] J. Jang et al., "Unifying vision-language representation space with single-tower transformer," in *Proc. Conf. Assoc. Adv. Artif. Intell.*, 2023, pp. 980–988.

[45] L. Yao et al., "DetCLIP: Dictionary-enriched visual-concept paralleled pre-training for open-world detection," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 9125–9138.

[46] H. Luo et al., "SegCLIP: Patch aggregation with learnable centers for open-vocabulary semantic segmentation," 2022, *arXiv:2211.14813*.

[47] S. Antol et al., "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2425–2433.

[48] A. Suhr et al., "A corpus for reasoning about natural language grounded in photographs," 2018, *arXiv: 1811.00491*.

[49] A. Karpathy et al., "Deep fragment embeddings for bidirectional image sentence mapping," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, vol. 27, pp. 1–9.

[50] F. Li et al., "Vision-language intelligence: Tasks, representation learning, and large models," 2022, *arXiv:2203.01922*.

[51] Y. Du et al., "A survey of vision-language pre-trained models," 2022, *arXiv:2202.10936*.

[52] F.-L. Chen et al., "VLP: A survey on vision-language pre-training," *Mach. Intell. Res.*, vol. 20, no. 1, pp. 38–56, 2023.

[53] P. Xu et al., "Multimodal learning with transformers: A survey," 2022, *arXiv:2206.06488*.

[54] X. Wang et al., "Large-scale multi-modal pre-trained models: A comprehensive survey," 2023, *arXiv:2302.10035*.

[55] S. Ren et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, vol. 28, pp. 1–9.

[56] L. -C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[57] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv: 2010.11929*.

[58] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 1–11.

[59] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[60] T. He et al., "Bag of tricks for image classification with convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 558–567.

[61] R. Zhang, "Making convolutional networks shift-invariant again," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7324–7334.

[62] N. Carion et al., "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[63] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.

[64] N. Mu et al., "Slip: Self-supervision meets language-image pre-training," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 529–544.

[65] J. Yang et al., "Unified contrastive learning in image-text-label space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19163–19173.

[66] K. He et al., "Masked autoencoders are scalable vision learners," 2021, *arXiv:2111.06377*.

[67] L. H. Li et al., "Grounded language-image pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10965–10975.

[68] A. V. D. Oord et al., "Representation learning with contrastive predictive coding," 2018, *arXiv: 1807.03748*.

[69] P. Khosla et al., "Supervised contrastive learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 18661–18673.

[70] H. Bao et al., "BEIT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.

[71] Z.-Y. Dou et al., "Coarse-to-fine vision-language pre-training with fusion in the backbone," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 32942–32956.

[72] H. Bao et al., "VLMO: Unified vision-language pre-training with mixture-of-modality-experts," 2021, *arXiv:2111.02358*.

[73] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2Text: Describing images using 1 million captioned photographs," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 1143–1151.

[74] X. Chen et al., "Microsoft COCO captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*.

[75] B. Thomee et al., "YFCC100M: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[76] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.

[77] P. Sharma et al., "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2556–2565.

[78] J. Pont-Tuset et al., "Connecting vision and language with localized narratives," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 647–664.

[79] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3557–3567.

[80] K. Srinivasan et al., "Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 2443–2449.

[81] K. Desai et al., "RedCaps: Web-curated image-text data created by the people, for the people," 2021, *arXiv:2111.11431*.

[82] J. Gu et al., "Wukong: 100 million large-scale chinese cross-modal pre-training dataset and a foundation framework," 2022, *arXiv:2202.06767*.

[83] X. Chen et al., "PaLI: A jointly-scaled multilingual language-image model," 2022, *arXiv:2209.06794*.

[84] J. Lee et al., "UniCLIP: Unified framework for contrastive language-image pre-training," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 1008–1019.

[85] S. Shao et al., "Objects365: A large-scale, high-quality dataset for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8429–8438.

[86] A. Kamath et al., "MDETR-modulated detection for end-to-end multi-modal understanding," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 1780–1790.

[87] M. Cao et al., "Image-text retrieval: A survey on recent research and development," 2022, *arXiv:2203.14713*.

[88] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[89] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshop*, 2004, pp. 178–178.

[90] M. Everingham et al., "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[91] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, 2008, pp. 722–729.

[92] Y. Netzer et al., "Reading digits in natural images with unsupervised feature learning," in *Proc. Int. Conf. Neural Inf. Process. Syst. Workshop*, 2011, pp. 1–9.

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

[93] A. Coates et al., "An analysis of single-layer networks in unsupervised feature learning," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 215–223.
[94] J. Stallkamp et al., "The German traffic sign recognition benchmark: A multi-class classification competition," in *Proc. Int. Joint Conf. Neural Netw.*, 2011, pp. 1453–1460.
[95] A. Mishra et al., "Scene text recognition using higher order language priors," in *Proc. Brit. Mach. Vis. Conf.*, 2012.
[96] S. Maji et al., "Fine-grained visual classification of aircraft," 2013, *arXiv:1306.5151*.
[97] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in *Proc. Neural Inf. Process.: 20th Int. Conf.*, 2013, pp. 117–124.
[98] R. Socher et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1631–1642.
[99] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3606–3613.
[100] T. Berg, J. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur, "Birdsnap: Large-scale fine-grained visual categorization of birds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2011–2018.
[101] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE Proc.*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[102] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1988–1997.
[103] B. S. Veeling et al., "Rotation equivariant CNNs for digital pathology," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2018, pp. 210–218.
[104] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.
[105] P. Young et al., "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Proc. Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, 2014.
[106] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[107] A. Gupta, P. Dollár, and R. Girshick, "LVIS: A dataset for large vocabulary instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5351–5359.
[108] C. Li et al., "Elevater: A benchmark and toolkit for evaluating language-augmented visual models," 2022, *arXiv:2204.08790*.
[109] R. Mottaghi et al., "The role of context for object detection and semantic segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 891–898.
[110] M. Cordts et al., "The CityScapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.
[111] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5122–5130.
[112] B. Wu et al., "Data efficient language-supervised zero-shot recognition with optimal transport distillation," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–13.
[113] Y. Li et al., "Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm," in *Proc. Int. Conf. Learn. Representations*, 2021.
[114] Q. Cui et al., "Contrastive vision-language pre-training with limited resources," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 236–253.
[115] L. Yuan et al., "Florence: A new foundation model for computer vision," 2021, *arXiv:2111.11432*.
[116] Y. Gao et al., "Pyramidclip: Hierarchical feature alignment for vision-language model pretraining," 2022, *arXiv:2204.14095*.
[117] A. Yang et al., "Chinese clip: Contrastive vision-language pretraining in chinese," 2022, *arXiv:2211.01335*.
[118] X. Zhai et al., "LiT: Zero-shot transfer with locked-image text tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18102–18112.
[119] Z. Chen et al., "Altclip: Altering the language encoder in clip for extended language capabilities," 2022, *arXiv:2211.06679*.
[120] B. Ko and G. Gu, "Large-scale bilingual language-image contrastive learning," 2022, *arXiv:2203.14463*.
[121] J. Zhou et al., "Non-contrastive learning meets language-image pre-training," 2022, *arXiv:2210.09304*.
[122] S. Shen et al., "K-lite: Learning transferable visual models with external knowledge," 2022, *arXiv:2204.09222*.
[123] R. Huang et al., "NLIP: Noise-robust language-image pre-training," 2022, *arXiv:2212.07086*.
[124] S. Geng et al., "HiCLIP: Contrastive language-image pretraining with hierarchy-aware attention," 2023, *arXiv:2303.02995*.
[125] C. -W. Xie, S. Sun, X. Xiong, Y. Zheng, D. Zhao, and J. Zhou, "RA-CLIP: Retrieval augmented contrastive language-image pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19265–19274.
[126] L. Fan et al., "Improving clip training with language rewrites," 2023, *arXiv:2305.20088*.
[127] K. Yang et al., "ALIP: Adaptive language-image pre-training with synthetic caption," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 2910–2919.
[128] X. Deng et al., "GrowCLIP: Data-aware automatic model growing for large-scale contrastive language-image pre-training," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 22121–22132.
[129] J. Xu et al., "GroupViT: Semantic segmentation emerges from text supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18113–18123.
[130] K. Ranasinghe et al., "Perceptual grouping in vision-language models," 2022, *arXiv:2210.09996*.
[131] Y. Zhong et al., "RegionCLIP: Region-based language-image pretraining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16772–16782.
[132] C. Ma et al., "Understanding and mitigating overfitting in prompt tuning for vision-language models," 2022, *arXiv:2211.02219*.
[133] A. Bulat and G. Tzimiropoulos, "Language-aware soft prompting for vision & language foundation models," 2022, *arXiv:2210.01115*.
[134] Y. Lu, J. Liu, Y. Zhang, Y. Liu, and X. Tian, "Prompt distribution learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5196–5205.
[135] Y. Lu, J. Liu, Y. Zhang, Y. Liu, and X. Tian, "Variational prompt tuning improves generalization of vision-language models," 2022, *arXiv:2210.02390*.
[136] B. Zhu et al., "Prompt-aligned gradient for prompt tuning," 2022, *arXiv:2205.14865*.
[137] X. He et al., "CPL: Counterfactual prompt learning for vision and language models," 2022, *arXiv:2210.10362*.
[138] G. Chen et al., "Prompt learning with optimal transport for vision-language models," 2022, *arXiv:2210.01253*.
[139] X. Sun et al., "DualCoOp: Fast adaptation to multi-label recognition with limited annotations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 30569–30582.
[140] Z. Guo et al., "Texts as images in prompt tuning for multi-label image recognition," 2022, *arXiv:2211.12739*.
[141] K. Ding et al., "Prompt tuning with soft context sharing for vision-language models," 2022, *arXiv:2208.13474*.
[142] Y. Rao et al., "DenseCLIP: Language-guided dense prediction with context-aware prompting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18061–18070.
[143] T. Huang et al., "Unsupervised prompt learning for vision-language models," 2022, *arXiv:2204.03649*.
[144] M. Shu et al., "Test-time prompt tuning for zero-shot generalization in vision-language models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 14274–14289.
[145] H. Yao, R. Zhang, and C. Xu, "Visual-language prompt tuning with knowledge-guided context optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6757–6767.
[146] T.-Y. Wu et al., "Protect: Prompt tuning for hierarchical consistency," 2023, *arXiv:2306.02240*.
[147] H. Bahng et al., "Exploring visual prompts for adapting large-scale models," 2022, *arXiv:2203.1727*.
[148] J. Rong et al., "Retrieval-enhanced visual prompt learning for few-shot classification," 2023, *arXiv:2306.02243*.
[149] Y. Zang et al., "Unified vision and language prompt learning," 2022, *arXiv:2210.07225*.
[150] S. Shen et al., "Multitask vision-language prompt tuning," 2022, *arXiv:2211.11720*.

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.

[151] M. U. Khattak et al., "Maple: Multi-modal prompt learning," 2022, *arXiv:2210.03117*.
[152] Y. Xing et al., "Class-aware visual prompt tuning for vision-language pre-trained model," 2022, *arXiv:2208.08340*.
[153] O. Pantazis et al., "SVL-adapter: Self-supervised adapter for vision-language pretrained models," 2022, *arXiv:2210.03794*.
[154] V. Udandarao et al., "SuS-X: Training-free name-only transfer of vision-language models," 2022, *arXiv:2211.16198*.
[155] J. Kahana et al., "Improving zero-shot models with label distribution priors," 2022, *arXiv:2212.00784*.
[156] F. Peng et al., "SgVA-CLIP: Semantic-guided visual adapting of vision-language models for few-shot image classification," 2022, *arXiv:2211.16191*.
[157] R. Zhang et al., "VT-CLIP: Enhancing vision-language models with visual-guided texts," 2021, *arXiv:2112.02399*.
[158] Z. Guo et al., "CALIP: Zero-shot enhancement of clip with parameter-free attention," 2022, *arXiv:2209.14169*.
[159] T. Yu, Z. Lu, X. Jin, Z. Chen, and X. Wang, "Task residual for tuning vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10899–10909.
[160] S. Pratt et al., "What does a platypus look like? Generating customized prompts for zero-shot image classification," 2022, *arXiv:2209.03320*.
[161] S. Menon and C. Vondrick, "Visual classification via description from large language models," 2022, *arXiv:2210.07183*.
[162] M. Wortsman et al., "Robust fine-tuning of zero-shot models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7949–7961.
[163] C. Zhou et al., "Extract free dense labels from clip," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 696–712.
[164] J. Li et al., "Masked unsupervised self-training for zero-shot image classification," 2022, *arXiv:2206.02967*.
[165] P. Liu et al., "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, 2023.
[166] M. Jia et al., "Visual prompt tuning," 2022, *arXiv:2203.12119*.
[167] N. Houlsby et al., "Parameter-efficient transfer learning for NLP," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2790–2799.
[168] R. Zhang et al., "PointCLIP: Point cloud understanding by CLIP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8542–8552.
[169] T. Huang et al., "CLIP2point: Transfer CLIP to point cloud classification with image-depth pre-training," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 22100–22110.
[170] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "Side adapter network for open-vocabulary semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2945–2954.
[171] G. Chen et al., "Tem-adapter: Adapting image-text pretraining for video question answer," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 13899–13909.
[172] T. Brown et al., "Language models are few-shot learners," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.
[173] Y. Zang et al., "Open-vocabulary DETR with conditional matching," 2022, *arXiv:2203.11876*.
[174] Z. Zhou et al., "ZegCLIP: Towards adapting clip for zero-shot semantic segmentation," 2022, *arXiv:2212.03588*.
[175] X. Zhou et al., "Detecting twenty-thousand classes using image-level supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 350–368.
[176] Z. Ma et al., "Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14054–14063.
[177] H. A. Rasheed et al., "Bridging the gap between object and image-level representations for open-vocabulary detection," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 33781–33794.
[178] J. Xie and S. Zheng, "ZSD-YOLO: Zero-shot YOLO detection using vision-language knowledgedistillation," 2021, *arXiv:2109.12066*.
[179] L. Wang et al., "Object-aware distillation pyramid for open-vocabulary object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11186–11196.
[180] S. Wu et al., "Aligning bag of regions for open-vocabulary object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15254–15264.
[181] D. Kim, A. Angelova, and W. Kuo, "Region-aware pretraining for open-vocabulary object detection with vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11144–11154.
[182] C. Feng et al., "PromptDet: Towards open-vocabulary detection using uncurated images," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 701–717.
[183] M. Gao et al., "Open vocabulary object detection with pseudo bounding-box labels," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 266–282.
[184] D. Huynh, J. Kuen, Z. Lin, J. Gu, and E. Elhamifar, "Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7010–7021.
[185] Y. Long et al., "P3OVD: Fine-grained visual-text prompt-driven self-training for open-vocabulary object detection," 2022, *arXiv:2211.00849*.
[186] G. Ghiasi et al., "Scaling open-vocabulary image segmentation with image-level labels," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 540–557.
[187] M. Xu et al., "A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 736–753.
[188] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7076–7086.
[189] B. Li et al., "Language-driven semantic segmentation," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–13.
[190] N. Zabari and Y. Hoshen, "Semantic segmentation in-the-wild without seeing any segmentation examples," 2021, *arXiv:2112.03185*.
[191] J. Qin et al., "FreeSeg: Unified, universal and open-vocabulary image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19446–19455.
[192] Y. Lin et al., "Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation," 2022, *arXiv:2212.09506*.
[193] J. Xie, X. Hou, K. Ye, and L. Shen, "CLIMS: Cross language image matching for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4483–4492.
[194] M. Cherti et al., "Reproducible scaling laws for contrastive language-image learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2818–2829.
[195] T. Wang and N. Li, "Learning to detect and segment for open vocabulary object detection," 2022, *arXiv:2212.12130*.
[196] C. Lin et al., "Learning object-language alignments for open-vocabulary object detection," 2022, *arXiv:2211.14843*.
[197] W. Kuo et al., "F-VLM: Open-vocabulary object detection upon frozen vision and language models," 2022, *arXiv:2209.15639*.
[198] M. Minderer et al., "Simple open-vocabulary object detection with vision transformers," 2022, *arXiv:2205.06230*.
[199] S. Zhao et al., "Exploiting unlabeled data with vision and language models for object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 159–175.
[200] Y. Xian, S. Choudhury, Y. He, B. Schiele, and Z. Akata, "Semantic projection network for zero-and few-label semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8256–8265.
[201] F. Liang et al., "Open-vocabulary semantic segmentation with mask-adapted CLIP," 2022, *arXiv:2210.04150*.
[202] G. Shin et al., "RECO: Retrieve and co-segment for zero-shot transfer," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 33754–33767.

Jingyi Zhang (Graduate Student Member, IEEE) received the BSc degree in electronic information science and technology from the University of Electronic Science and Technology of China (UESTC) and the MSc degree in signal processing from Nanyang Technological University (NTU). She is currently a research associate and working toward the PhD degree with School of Computer Science and Engineering, NTU. Her research interests include computer vision, object detection.

Jiaxing Huang (Graduate Student Member, IEEE) received the BEng and MSc degrees in EEE from the University of Glasgow, UK, and the Nanyang Technological University (NTU), Singapore, respectively. He is currently a research associate and working toward the PhD degree with School of Computer Science and Engineering, NTU, Singapore. His research interests include computer vision and machine learning.

Sheng Jin received the BSc degree in applied mathematics from the Harbin Institute of Technology and the PhD degree in computer science and technology Om Harbin Institute of Technology. He is currently a research fellow with Nanyang Technology University (NTU), Singapore. His research interests include computer vision and machine learning.

Shijian Lu received the PhD degree in electrical and computer engineering from the National University of Singapore. He is an associate professor with the School of Computer Science and Engineering with the Nanyang Technological University, Singapore. His major research interests include image and video analytics, visual intelligence, and machine learning.

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 03:03:18 UTC from IEEE Xplore. Restrictions apply.