Vision-Language Models for Vision Tasks: A Survey

Jingyi Zhang†, Jiaxing Huang†, Sheng Jin and Shijian Lu*

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

Index Terms—Visual recognition, vision-language model, pre-training, transfer learning, knowledge distillation, image classification, object detection, semantic segmentation, deep neural network, deep learning, big model, big data

1 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.

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 Figs. 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.

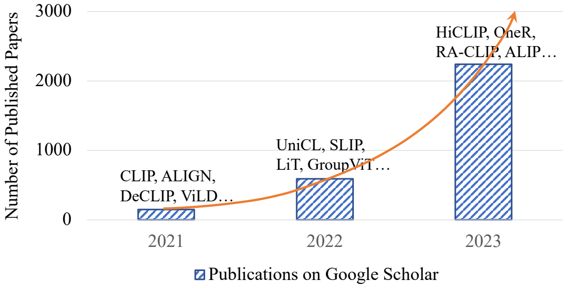


Fig. 1: Number of publications on visual recognition VLMs (from Google Scholar). The publications grow exponentially since the pioneer study CLIP [10] in 2021.

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 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

- *All authors are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore.*
- *† denotes equal contribution; * denotes corresponding author.*

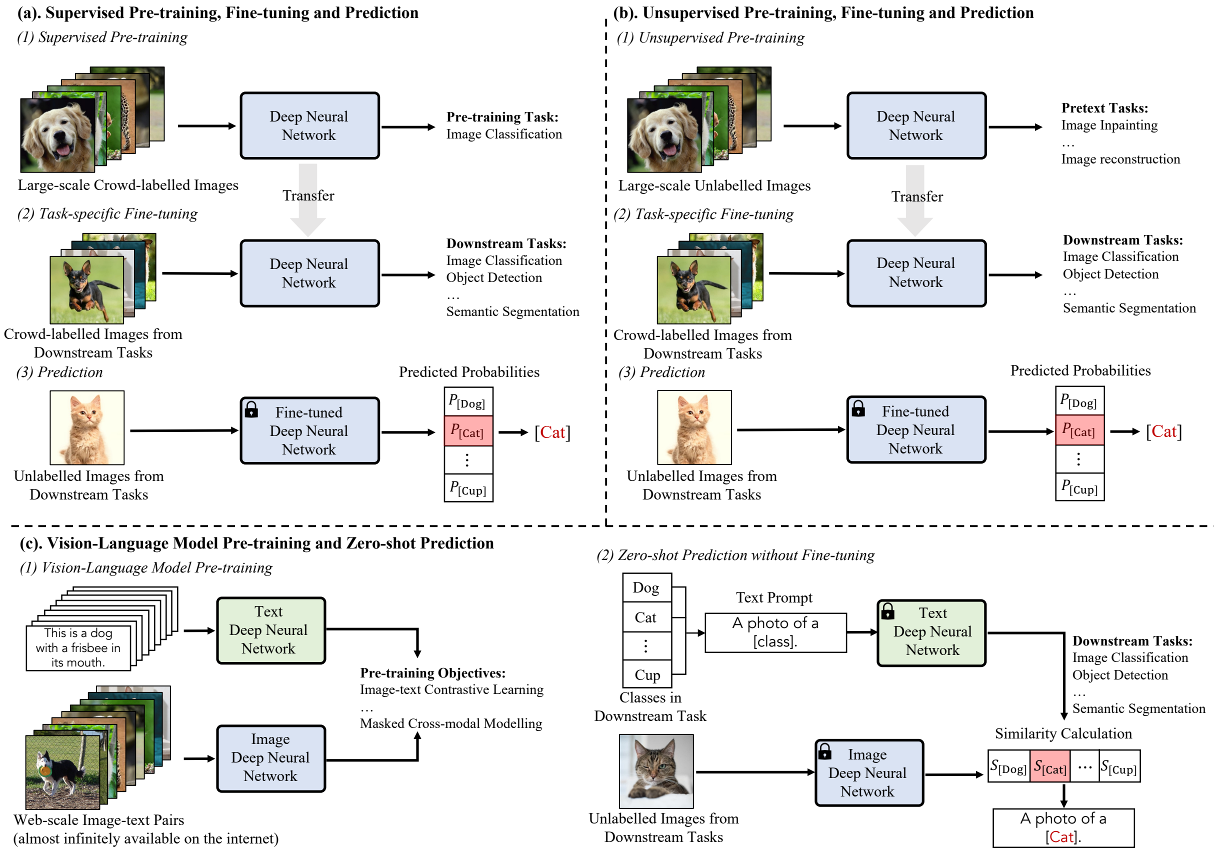

(a). Supervised Pre-training, Fine-tuning and Prediction
(1) Supervised Pre-training
Large-scale Crowd-labelled Images
Deep Neural Network
Pre-training Task:
Image Classification
Transfer
(2) Task-specific Fine-tuning
Crowd-labelled Images from Downstream Tasks
Deep Neural Network
Downstream Tasks:
Image Classification
Object Detection
…
Semantic Segmentation
(3) Prediction
Unlabelled Images from Downstream Tasks
Fine-tuned Deep Neural Network
Predicted Probabilities
$P_{[Dog]}$
$P_{[Cat]}$
⋮
$P_{[Cup]}$
[Cat]
(b). Unsupervised Pre-training, Fine-tuning and Prediction
(1) Unsupervised Pre-training
Large-scale Unlabelled Images
Deep Neural Network
Pretext Tasks:
Image Inpainting
…
Image reconstruction
Transfer
(2) Task-specific Fine-tuning
Crowd-labelled Images from Downstream Tasks
Deep Neural Network
Downstream Tasks:
Image Classification
Object Detection
…
Semantic Segmentation
(3) Prediction
Unlabelled Images from Downstream Tasks
Fine-tuned Deep Neural Network
Predicted Probabilities
$P_{[Dog]}$
$P_{[Cat]}$
⋮
$P_{[Cup]}$
[Cat]
(c). Vision-Language Model Pre-training and Zero-shot Prediction
(1) Vision-Language Model Pre-training
This is a dog with a frisbee in its mouth.
Text Deep Neural Network
Pre-training Objectives:
Image-text Contrastive Learning
…
Masked Cross-modal Modelling
Image Deep Neural Network
Web-scale Image-text Pairs
(almost infinitely available on the internet)
(2) Zero-shot Prediction without Fine-tuning
Dog
Cat
⋮
Cup
Classes in Downstream Task
Text Prompt
A photo of a [class].
Text Deep Neural Network
Downstream Tasks:
Image Classification
Object Detection
…
Semantic Segmentation
Similarity Calculation
$S_{[Dog]}$
$S_{[Cat]}$
…
$S_{[Cup]}$
A photo of a [Cat].
Image Deep Neural Network
Unlabelled Images from Downstream Tasks

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.

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 2 introduces the paradigm development of visual recognition and several related surveys. Section 3 describes the foundations of VLMs, including widely used deep network architectures, pre-training objectives, pre-training frameworks and downstream tasks in VLM evaluations. Section 4 introduces the commonly used datasets in VLM pre-training and evaluations. Section 5 reviews and categorizes VLM pre-training methods. Sections 6 and 7 provide a systematic review of transfer learning and knowledge distillation approaches for VLMs , respectively. Section 8 benchmarks the reviewed methods on multiple widely-adopted datasets. Finally, we share several promising VLM research directions in Section 9.

2 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.

2.1 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.

2.1.1 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.

2.1.2 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.

2.1.3 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.

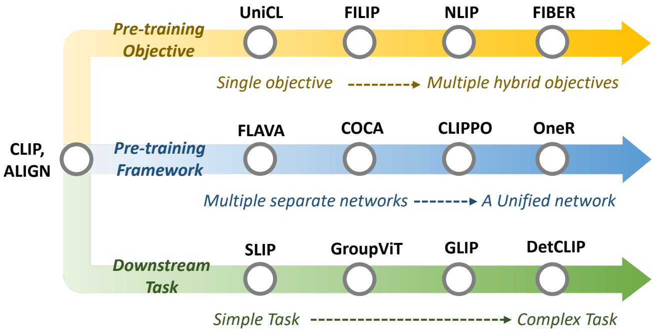


Fig. 3: Illustration of development of VLMs for visual recognition.

2.1.4 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].

2.1.5 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

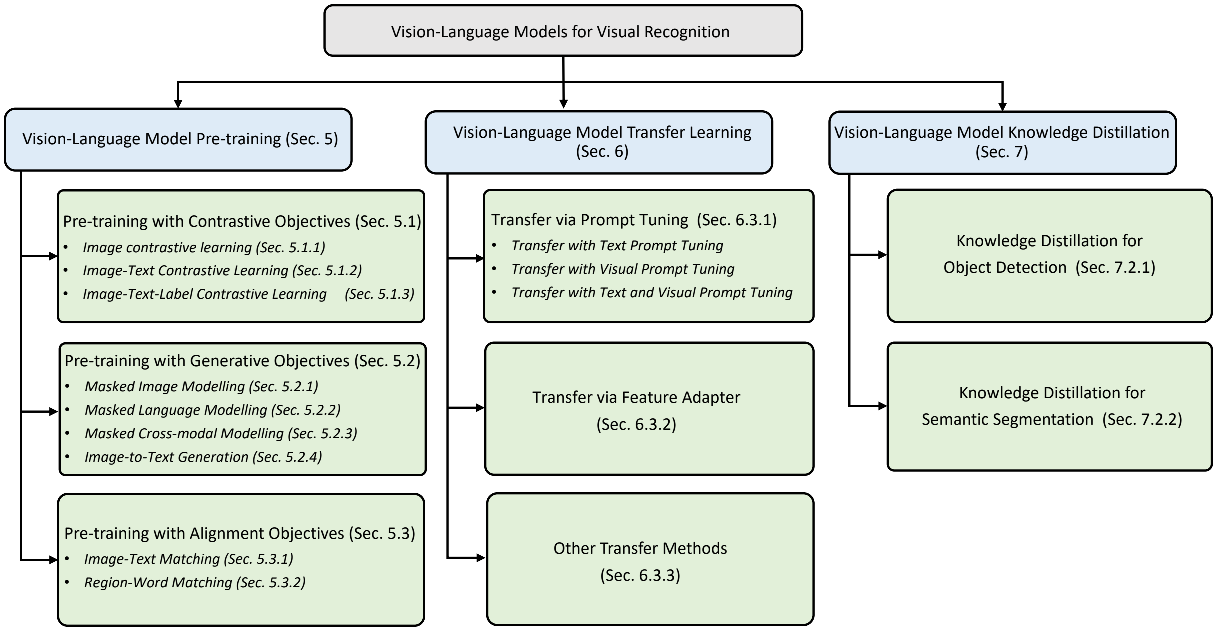

Vision-Language Models for Visual Recognition
Vision-Language Model Pre-training (Sec. 5)
Vision-Language Model Transfer Learning (Sec. 6)
Vision-Language Model Knowledge Distillation (Sec. 7)
Pre-training with Contrastive Objectives (Sec. 5.1)
Image contrastive learning (Sec. 5.1.1)
Image-Text Contrastive Learning (Sec. 5.1.2)
Image-Text-Label Contrastive Learning (Sec. 5.1.3)
Pre-training with Generative Objectives (Sec. 5.2)
Masked Image Modelling (Sec. 5.2.1)
Masked Language Modelling (Sec. 5.2.2)
Masked Cross-modal Modelling (Sec. 5.2.3)
Image-to-Text Generation (Sec. 5.2.4)
Pre-training with Alignment Objectives (Sec. 5.3)
Image-Text Matching (Sec. 5.3.1)
Region-Word Matching (Sec. 5.3.2)
Transfer via Prompt Tuning (Sec. 6.3.1)
Transfer with Text Prompt Tuning
Transfer with Visual Prompt Tuning
Transfer with Text and Visual Prompt Tuning
Transfer via Feature Adapter (Sec. 6.3.2)
Other Transfer Methods (Sec. 6.3.3)
Knowledge Distillation for Object Detection (Sec. 7.2.1)
Knowledge Distillation for Semantic Segmentation (Sec. 7.2.2)

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

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.

2.2 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] 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.

2.3 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, 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.

3 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.

3.1 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.

3.1.1 *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.

3.1.2 *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.

3.2 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.

3.2.1 *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.*, $\mathcal{L}_{\text{infoNCE}}^{IT} = \mathcal{L}_{I\rightarrow T} + \mathcal{L}_{T\rightarrow I}$, where $\mathcal{L}_{I\rightarrow T}$ contrasts the query image with the text keys while $\mathcal{L}_{T\rightarrow I}$ contrasts the query text with image keys. Given a batch of B image-text pairs, $\mathcal{L}_{I\rightarrow T}$ and $\mathcal{L}_{T\rightarrow I}$ are defined as follows:

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

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

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 Eqs. 2 and 3 as follows:

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

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

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 Eqs. 4 and 5, the image-text-label infoNCE loss is defined as: $\mathcal{L}_{\text{infoNCE}}^{ITL} = \mathcal{L}_{I\rightarrow T}^{ITL} + \mathcal{L}_{T\rightarrow I}^{ITL}$.

3.2.2 *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.

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.

3.2.3 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:

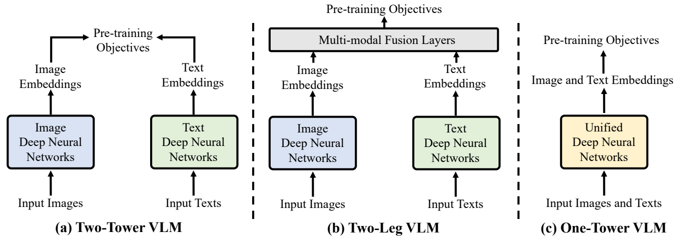
$$\mathcal{L}_{RW} = p\log\mathcal{S}^r(r^I,w^T) + (1-p)\log(1-\mathcal{S}^r(r^I,w^T)), \tag{11}$$

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".

3.3 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.



(a) Two-Tower VLM (b) Two-Leg VLM (c) One-Tower VLM

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

3.4 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.

3.4.1 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.

3.4.2 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

TABLE 1: Summary of the widely used image-text datasets for VLM pre-training. [link] directs to dataset websites.

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	✗

action recognition [28], [29] have been widely adopted in such evaluations, where video clips are often sub-sampled for efficient recognition in action recognition tasks [10].

4 Datasets

This section summarizes the commonly used datasets for VLM pre-training and evaluations, as detailed in Tables 1-2.

4.1 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], [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.

4.2 Datasets for VLM Evaluation

Many datasets have been adopted in VLM evaluations as shown in Table 2, 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). 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].

5 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 3.

5.1 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].

5.1.1 *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 Eq. 1 for learning discriminative image features.

5.1.2 *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 Eq. 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.

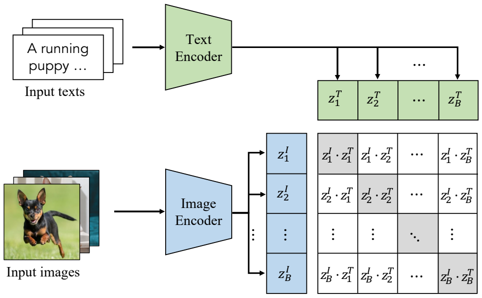


Fig. 6: Illustration of the image-text contrastive learning in CLIP [10]. Figure is reproduced from [10].

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.

TABLE 2: Summary of the widely-used visual recognition datasets for VLM evaluation. [link] directs to dataset websites.

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

For example, DeCLIP [113] introduces nearest-neighbor supervision to utilize the information from similar pairs, enabling 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. PyramidCLIP [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.

5.1.3 Image-Text-Label Contrastive Learning

This type of pre-training introduces image classification labels [65] into the image-text contrast as defined in Eq. 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 with around 900M image-text pairs, leading to outstanding performance in various downstream recognition tasks.

5.1.4 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 Sec. 3.2.1.

TABLE 3: Summary of vision-language model pre-training methods. 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.

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.

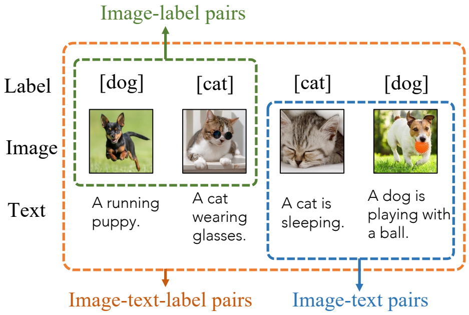


Fig. 7: Illustration of the image-text-label space proposed in UniCL [65]. Figure is reproduced from [65].

5.2 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.

5.2.1 Masked Image Modelling

This pre-training objective guides to learn image context information by masking and reconstructing images as defined in Eq. 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.

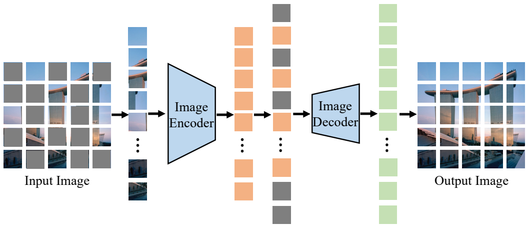


Fig. 8: Illustration of masked image modelling [66]. Figure is reproduced from [66].

5.2.2 Masked Language Modelling

Masked language modelling, a widely-adopted pre-training objective in NLP as defined in Eq. 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 masked language modelling [14] as one of the VLM pre-training objectives to extract better language features.

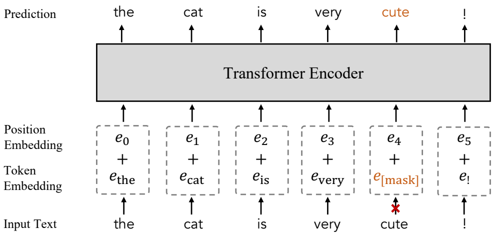


Fig. 9: Illustration of masked language modelling [14].

5.2.3 Masked Cross-Modal Modelling

Masked cross-modal modelling masks and reconstructs both image patches and text tokens jointly as defined in Eq. 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.

5.2.4 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 Eq. 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.

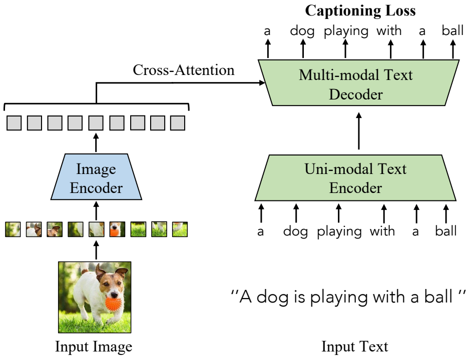


Fig. 10: A simplified illustration of image-to-caption generation in COCA [19]. Figure is reproduced based on [19].

5.2.5 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].

5.3 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.

5.3.1 Image-Text Matching

Image-text matching models global image-text correlation by directly aligning paired images and texts as defined in Eq. 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.

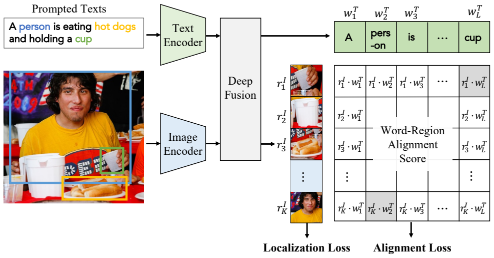

Prompted Texts
A person is eating hot dogs and holding a cup
Text Encoder
Image Encoder
Deep Fusion
w_1^T w_2^T w_3^T w_L^T
A
pers -on
is
…
cup
r_1^I
r_2^I
r_3^I
r_K^I
$r_1^I \cdot w_1^T$ $r_1^I \cdot w_2^T$ $r_1^I \cdot w_3^T$ … $r_1^I \cdot w_L^T$
$r_2^I \cdot w_1^T$ $r_2^I \cdot w_L^T$
$r_3^I \cdot w_1^T$ $r_3^I \cdot w_L^T$
$r_K^I \cdot w_1^T$ $r_K^I \cdot w_2^T$ $r_K^I \cdot w_3^T$ … $r_K^I \cdot w_L^T$
Word-Region Alignment Score
Localization Loss
Alignment Loss

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

5.3.2 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.

5.3.3 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].

5.4 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

TABLE 4: Summary of VLM transfer learning methods. 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.

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.

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.

6 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.

6.1 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.

6.2 Common Setup of Transfer Learning

Three transfer setups have been explored for mitigating the domain gaps described in Sec. 6.1, 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.

6.3 Common Transfer Learning Methods

As shown in Table 4, we broadly group existing VLM transfer methods into three categories including prompt tuning approaches, feature adapter approaches, and others.

6.3.1 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.

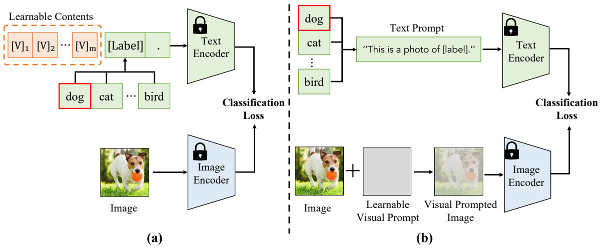


Fig. 12: Illustration of text prompt learning [31] in (a) and visual prompt learning [147] in (b).

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 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.

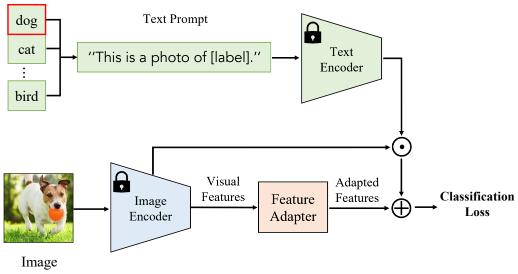


Fig. 13: Illustration of feature adapter [33].

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].

6.3.2 *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.

6.3.3 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.

6.4 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 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.

7 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.

7.1 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].

7.2 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 5 shows a list of VLM knowledge distillation methods.

7.2.1 Knowledge Distillation for Object Detection

Open-vocabulary object detection [193] 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 [186] explores hierarchical global-local knowledge distillation, and RKD [187] explores region-based knowledge distillation for better aligning region-level and image-level embeddings. ZSD-YOLO [198] introduces self-labelling data augmentation for exploiting CLIP for better object detection. OADP [201] preserves proposal features while transferring contextual knowledge. BARON [200] uses neighborhood sampling to distill a bag of regions instead of individual regions. RO-ViT [199] 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 [188] introduces regional prompt learning for aligning word embeddings with regional image embeddings. Additionally, several studies [180], [181], [189], [194], [197] explore VLM-predicted pseudo labels to improve object detectors. For example, PB-OVD [189] trains object detectors with VLM-predicted pseudo bounding boxes while XPM [194] introduces a robust cross-modal pseudo-labeling strategy that employs VLM-generated pseudo masks for open-vocabulary instance segmentation. P^{3}OVD [197] exploits prompt-driven self-training that refines the VLM-generated pseudo labels with fine-grained prompt tuning.

7.2.2 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], [180], [181] achieve open-vocabulary semantic segmentation by first class-agnostic segmentation by grouping pixels into multiple segments and then segment recognition with CLIP. CLIPSeg [175] introduces a lightweight transformer decoder to extend CLIP for semantic segmentation. LSeg [176] 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 [177] distill

TABLE 5: Summary of VLM knowledge distillation methods. [code] directs to code websites.

Task	Method	Contribution
Semantic Segmentation	CLIPSeg [175] [code]	Extend CLIP by introducing a lightweight transformer-based decoder.
	ZegFormer [35] [code]	Group the pixels into segments and preforms zero-shot classification task on the segments.
	LSeg [176] [code]	Propose language-driven semantic segmentation by matching pixel and text embeddings.
	SSIW [177]	Introduce a test-time augmentation technique to refine the pseudo labels generated by CLIP.
	MaskClip+ [163] [code]	Perform self-training with the pseudo labels generated by MaskClip (modified from CLIP).
	ZegClip [174] [code]	Propose deep prompt tuning, non-mutually exclusive loss and relationship descriptor.
	Fusioner [178] [code]	Introduce cross-modality fusion that aligns the visual representation with language concept.
	OVSeg [179] [code]	Adapt CLIP with the region-word pairs generated by the modified MaskFormer.
	ZSSeg [180] [code]	Propose to first generate mask proposals and then classifies the generated mask proposals.
	OpenSeg [181] [code]	Propose to align each word in the caption with the generated segmentation masks.
	ReCo [182] [code]	Propose language-guided co-segmentation with the CLIP-retrieved images.
	CLIMS [183] [code]	Use CLIP to generate high-quality class activation maps w/o involving irrelevant background.
	CLIP-ES [184] [code]	Employ CLIP to refine the class activation map for weakly-supervised segmentation.
	FreeSeg [185] [code]	Propose a unified, universal and open-Vocabulary image segmentation network.
Object Detection	ViLD [36] [code]	Propose to distill knowledge from a pre-trained VLM into a two-stage object detector.
	DetPro [37] [code]	Propose to learn continuous prompt representations for open-vocabulary object detection.
	HierKD [186] [code]	Propose hierarchical knowledge distillation for global-level and instance-level distillation.
	RKD [187] [code]	Propose region-based knowledge distillation for aligning region- and image-level embeddings.
	PromptDet [188] [code]	Introduce regional prompting for aligning text embeddings with regional image embeddings.
	PB-OVD [189] [code]	Propose to train object detectors with the pseudo bounding-box labels generated by VLMs.
	CondHead [190]	Propose semantic-visual alignment for better box regression and mask segmentation.
	VLDet [191] [code]	Achieve open-vocabulary object detection by the bipartite matching between regions and words.
	F-VLM [192]	Propose to simply build a detection head upon the pre-trained VLM for object localization.
	OV-DETR [173] [code]	Achieve open-vocabulary detection transformer with a binary matching strategy.
	Detic [193] [code]	Enlarge detection vocabulary using image-level supervision and pre-trained CLIP text encoder.
	XPM [194] [code]	Design cross-modal pseudo-labeling to let VLMs generate caption-driven pseudo masks.
	OWL-ViT [195] [code]	Propose ViT-based open-vocabulary detector by adding object classification/localization head.
	VL-PLM [196] [code]	Leverage VLMs for assigning category labels to the generated pseudo bounding boxes.
	P^{3}OVD [197]	Propose prompt-driven self-training that refines the pseudo labels generated by VLMs.
	ZSD-YOLO [198] [code]	Leverage CLIP for object detection with a self-labeling based data augmentation techiqnuie.
	RO-ViT [199]	Bridge the gap of VLM pre-training and downstream open-vocabulary detection.
	BARON [200] [code]	Propose neighborhood sampling strategy to align the embedding of bag of regions.
	OADP [201] [code]	Propose object-aware distillation network to preserve and transfer contextual knowledge.

knowledge with VLM-predicted pixel-level pseudo labels. FreeSeg [185] generates mask proposals firstly 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 [184] 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 [183] employs CLIP knowledge to generate high-quality class activation maps for better weakly-supervised semantic segmentation.

7.3 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.

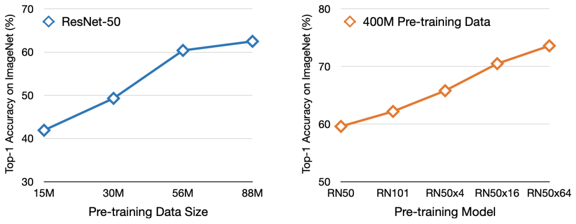


Fig. 14: Performance versus data size and model size. It shows that scaling up either the pre-training data [113] or the pre-training model [10] benefits VLM consistently.

8 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 5-7.

8.1 Performance of VLM Pre-training

As discussed in Sec. 3.4, *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 6 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 6 as well as Fig. 14: 1) VLM performance is usually up to the size of

TABLE 6: 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

TABLE 7: 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 8: 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	-

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 6 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 7 and 8 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 7 and 8 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.

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], [202]; (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.

8.2 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 9 shows the results on 11 widely adopted image classification datasets (*e.g.*,

TABLE 9: 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	-

EuroSAT [104], UCF101 [29]) with different backbones such as CNN backbone ResNet-50 and Transformer backbones ViT-B and ViT-L. Note Table 9 summarizes the performance of 16-shot setup for all *few-shot supervised* methods.

Three conclusions can be drawn from Table 9. 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.

8.3 Performance of VLM Knowledge Distillation

This section presents how VLM knowledge distillation helps in the tasks of object detection and semantic segmentation. Tables 10 and 11 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.

TABLE 10: Performance of VLM knowledge distillation on object detection. CLIP Transformer is CLIP text encoder.

Method	Vision-Language	COCO [106]			LVIS [107]			
	Model	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 [186]	CLIP ViT-B/32	53.5	27.3	-	-	-	-	-
RKD [187]	CLIP ViT-B/32	56.6	36.9	51.0	21.1	25.0	29.1	25.9
PromptDet [188]	CLIP Transformer	-	26.6	50.6	21.4	23.3	29.3	25.3
PB-OVD [189]	CLIP Transformer	46.1	30.8	42.1	-	-	-	-
CondHead [190]	CLIP ViT-B/32	60.8	29.8	49.0	18.8	28.3	33.7	28.8
VLDet [191]	CLIP Transformer	50.6	32.0	45.8	26.3	39.4	41.9	38.1
F-VLM [192]	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 [193]	CLIP Transformer	45.0	27.8	47.1	17.8	26.3	31.6	26.8
OWL-ViT [195]	CLIP ViT-B/32	-	-	28.1	18.9	-	-	22.1
VL-PLM [196]	CLIP ViT-B/32	60.2	34.4	53.5	-	-	-	22.2
P^3OVD [197]	CLIP ResNet-50	51.9	31.5	46.6	-	-	-	10.6
RO-ViT [199]	CLIP ViT-L/16	-	33.0	47.7	32.1	-	-	34.0
BARON [200]	CLIP ResNet-50	54.9	42.7	51.7	23.2	29.3	32.5	29.5
OADP [201]	CLIP ViT-B/32	53.3	30.0	47.2	21.9	28.4	32.0	28.7

8.4 Summary

Several conclusions can be drawn from Tables 6-11. 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

TABLE 11: 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 [203]	-	-	-	-	24.3	18.3	-
LSeg [35]	CLIP ResNet-101	-	-	-	-	47.4	-
ZegFormer [176]	CLIP ResNet-50	-	-	16.4	-	80.7	-
OVSeg [179]	CLIP Swin-B	9.0	12.4	29.6	55.7	94.5	-
ZSSeg [180]	CLIP ResNet-101	7.0	-	20.5	47.7	-	34.5
OpenSeg [181]	CLIP Eff-B7	6.3	9.0	21.1	42.1	-	-
ReCo [182]	CLIP ResNet-101	-	-	-	-	-	24.2
FreeSeg [185]	CLIP ViT-B/16	-	-	39.8	-	86.9	-

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.

9 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], 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 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.

10 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 pretraining which will greatly benefit the future research along this emerging but very promising research direction.

References

[1] A. Geiger et al. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pp. 3354–3361. IEEE, 2012.

[2] G. Cheng et al. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *JSTARS*, 13:3735–3756, 2020.

[3] H. A. Pierson and M. S. Gashler. Deep learning in robotics: a review of recent research. *Advanced Robotics*, 31(16):821–835, 2017.

[4] A. Krizhevsky et al. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

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

[6] K. He et al. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

[7] L. E. Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[8] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[9] A. Mathur and G. M. Foody. Multiclass and binary svm classification: Implications for training and classification users. *IEEE Geoscience and remote sensing letters*, 5(2):241–245, 2008.

[10] A. Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PMLR, 2021.

[11] R. Girshick. Fast r-cnn. In *ICCV*, pp. 1440–1448, 2015.

[12] K. He et al. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pp. 9729–9738, 2020.

[13] T. Chen et al. A simple framework for contrastive learning of visual representations. In *ICML*, pp. 1597–1607. PMLR, 2020.

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

[15] A. Radford et al. Improving language understanding by generative pre-training. 2018.

[16] A. Radford et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[17] C. Jia et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pp. 4904–4916. PMLR, 2021.

[18] L. Yao et al. Filip: Fine-grained interactive language-image pretraining. In *ICLR*, 2021.

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

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

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

[22] L. Bossard et al. Food-101–mining discriminative components with random forests. In *ECCV*, pp. 446–461. Springer, 2014.

[23] A. Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

[24] J. Xiao et al. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

[25] J. Krause et al. Collecting a large-scale dataset of fine-grained cars. 2013.

[26] O. M. Parkhi et al. Cats and dogs. In *CVPR*, pp. 3498–3505. IEEE, 2012.

[27] D. Kiela et al. The hateful memes challenge: Detecting hate speech in multimodal memes. *NeurIPS*, 33:2611–2624, 2020.

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

[29] K. Soomro et al. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

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

[31] K. Zhou et al. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022.

[32] K. Zhou et al. Conditional prompt learning for vision-language models. In *CVPR*, pp. 16816–16825, 2022.

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

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

[35] J. Ding et al. Decoupling zero-shot semantic segmentation. In *CVPR*, pp. 11583–11592, 2022.

[36] X. Gu et al. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2021.

[37] Y. Du et al. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, pp. 14084–14093, 2022.

[38] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004.

[39] L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[40] J. Deng et al. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. Ieee, 2009.

[41] K. He et al. Masked autoencoders are scalable vision learners. In *CVPR*, pp. 16000–16009, 2022.

[42] A. Singh et al. Flava: A foundational language and vision alignment model. In *CVPR*, pp. 15638–15650, 2022.

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

[44] J. Jang et al. Unifying vision-language representation space with single-tower transformer. In *AAAI*, volume 37, pp. 980–988, 2023.

[45] L. Yao et al. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In *NeurIPS*.

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

[47] S. Antol et al. Vqa: Visual question answering. In *ICCV*, pp. 2425–2433, 2015.

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

[49] A. Karpathy et al. Deep fragment embeddings for bidirectional image sentence mapping. *NeurIPS*, 27, 2014.

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

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

[52] F.-L. Chen et al. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023.

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

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

[55] S. Ren et al. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015.

[56] L.-C. Chen et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2017.

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

[58] A. Vaswani et al. Attention is all you need. *NeurIPS*, 30, 2017.

[59] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pp. 6105–6114. PMLR, 2019.

[60] T. He et al. Bag of tricks for image classification with convolutional neural networks. In *CVPR*, pp. 558–567, 2019.

[61] R. Zhang. Making convolutional networks shift-invariant again. In *ICML*, pp. 7324–7334. PMLR, 2019.

[62] N. Carion et al. End-to-end object detection with transformers. In *ECCV*, pp. 213–229. Springer, 2020.

[63] E. Xie et al. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 34:12077–12090, 2021.

[64] N. Mu et al. Slip: Self-supervision meets language-image pre-training. In *ECCV*, pp. 529–544. Springer, 2022.

[65] J. Yang et al. Unified contrastive learning in image-text-label space. In *CVPR*, pp. 19163–19173, 2022.

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

[67] L. H. Li et al. Grounded language-image pre-training. In *CVPR*, pp. 10965–10975, 2022.

[68] A. v. d. Oord et al. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[69] P. Khosla et al. Supervised contrastive learning. *NeurIPS*, 33:18661–18673, 2020.

[70] H. Bao et al. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[71] Z.-Y. Dou et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *NeurIPS*.

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

[73] V. Ordonez et al. Im2text: Describing images using 1 million captioned photographs. *NeurIPS*, 24, 2011.

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

[75] B. Thomee et al. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[76] R. Krishna et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

[77] P. Sharma et al. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pp. 2556–2565, 2018.

[78] J. Pont-Tuset et al. Connecting vision and language with localized narratives. In *ECCV*, pp. 647–664. Springer, 2020.

[79] S. Changpinyo et al. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pp. 3558–3568, 2021.

[80] K. Srinivasan et al. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *ACM SIGIR*, pp. 2443–2449, 2021.

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

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

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

[84] J. Lee et al. Uniclip: Unified framework for contrastive language-image pre-training. In *NeurIPS*.

[85] S. Shao et al. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pp. 8430–8439, 2019.

[86] A. Kamath et al. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, pp. 1780–1790, 2021.

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

[88] Y. LeCun et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[89] L. Fei-Fei et al. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR workshop*, pp. 178–178. IEEE, 2004.

[90] M. Everingham et al. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[91] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pp. 722–729. IEEE, 2008.

[92] Y. Netzer et al. Reading digits in natural images with unsupervised feature learning. 2011.

[93] A. Coates et al. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

[94] J. Stallkamp et al. The german traffic sign recognition benchmark: a multi-class classification competition. In *IJCNN*, pp. 1453–1460. IEEE, 2011.

[95] A. Mishra et al. Scene text recognition using higher order language priors. In *BMVC-British machine vision conference*. BMVA, 2012.

[96] S. Maji et al. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[97] I. J. Goodfellow et al. Challenges in representation learning: A report on three machine learning contests. In *ICONIP*, pp. 117–124. Springer, 2013.

[98] R. Socher et al. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, pp. 1631–1642, 2013.

[99] M. Cimpoi et al. Describing textures in the wild. In *CVPR*, pp. 3606–3613, 2014.

[100] T. Berg et al. Birdsnap: Large-scale fine-grained visual categorization of birds. In *CVPR*, pp. 2011–2018, 2014.

[101] G. Cheng et al. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

[102] J. Johnson et al. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pp. 2901–2910, 2017.

[103] B. S. Veeling et al. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 210–218. Springer, 2018.

[104] P. Helber et al. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *JSTARS*, 12(7):2217–2226, 2019.

[105] P. Young et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.

[106] T.-Y. Lin et al. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755. Springer, 2014.

[107] A. Gupta et al. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, pp. 5356–5364, 2019.

[108] C. Li et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *arXiv preprint arXiv:2204.08790*, 2022.

[109] R. Mottaghi et al. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, pp. 891–898, 2014.

[110] M. Cordts et al. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pp. 3213–3223, 2016.

[111] B. Zhou et al. Scene parsing through ade20k dataset. In *CVPR*, pp. 633–641, 2017.

[112] B. Wu et al. Data efficient language-supervised zero-shot recognition with optimal transport distillation. In *ICLR*, 2021.

[113] Y. Li et al. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2021.

[114] Q. Cui et al. Contrastive vision-language pre-training with limited resources. In *ECCV*, pp. 236–253. Springer, 2022.

[115] L. Yuan et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[116] Y. Gao et al. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *arXiv preprint arXiv:2204.14095*, 2022.

[117] A. Yang et al. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022.

[118] X. Zhai et al. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, pp. 18123–18133, 2022.

[119] Z. Chen et al. Altclip: Altering the language encoder in clip for extended language capabilities. *arXiv preprint arXiv:2211.06679*, 2022.

[120] B. Ko and G. Gu. Large-scale bilingual language-image contrastive learning. *arXiv preprint arXiv:2203.14463*, 2022.

[121] J. Zhou et al. Non-contrastive learning meets language-image pre-training. *arXiv preprint arXiv:2210.09304*, 2022.

[122] S. Shen et al. K-lite: Learning transferable visual models with external knowledge. *arXiv preprint arXiv:2204.09222*, 2022.

[123] R. Huang et al. Nlip: Noise-robust language-image pre-training. *arXiv preprint arXiv:2212.07086*, 2022.

[124] S. Geng et al. Hiclip: Contrastive language-image pretraining with hierarchy-aware attention. *arXiv preprint arXiv:2303.02995*, 2023.

[125] C.-W. Xie et al. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *CVPR*, pp. 19265–19274, 2023.

[126] L. Fan et al. Improving clip training with language rewrites. *arXiv preprint arXiv:2305.20088*, 2023.

[127] K. Yang et al. Alip: Adaptive language-image pre-training with synthetic caption. In *ICCV*, pp. 2922–2931, 2023.

[128] X. Deng et al. Growclip: Data-aware automatic model growing for large-scale contrastive language-image pre-training. In *ICCV*, pp. 22178–22189, 2023.

[129] J. Xu et al. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, pp. 18134–18144, 2022.

[130] K. Ranasinghe et al. Perceptual grouping in vision-language models. *arXiv preprint arXiv:2210.09996*, 2022.

[131] Y. Zhong et al. Regionclip: Region-based language-image pretraining. In *CVPR*, pp. 16793–16803, 2022.

[132] C. Ma et al. Understanding and mitigating overfitting in prompt tuning for vision-language models. *arXiv preprint arXiv:2211.02219*, 2022.

[133] A. Bulat and G. Tzimiropoulos. Language-aware soft prompting for vision & language foundation models. *arXiv preprint arXiv:2210.01115*, 2022.

[134] Y. Lu et al. Prompt distribution learning. In *CVPR*, pp. 5206–5215, 2022.

[135] M. M. Derakhshani et al. Variational prompt tuning improves generalization of vision-language models. *arXiv preprint arXiv:2210.02390*, 2022.

[136] B. Zhu et al. Prompt-aligned gradient for prompt tuning. *arXiv preprint arXiv:2205.14865*, 2022.

[137] X. He et al. Cpl: Counterfactual prompt learning for vision and language models. *arXiv preprint arXiv:2210.10362*, 2022.

[138] G. Chen et al. Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022.

[139] X. Sun et al. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. In *NeurIPS*.

[140] Z. Guo et al. Texts as images in prompt tuning for multi-label image recognition. *arXiv preprint arXiv:2211.12739*, 2022.

[141] K. Ding et al. Prompt tuning with soft context sharing for vision-language models. *arXiv preprint arXiv:2208.13474*, 2022.

[142] Y. Rao et al. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, pp. 18082–18091, 2022.

[143] T. Huang et al. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.

[144] M. Shu et al. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*.

[145] H. Yao et al. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, pp. 6757–6767, 2023.

[146] T.-Y. Wu et al. Protect: Prompt tuning for hierarchical consistency. *arXiv preprint arXiv:2306.02240*, 2023.

[147] H. Bahng et al. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 1(3):4, 2022.

[148] J. Rong et al. Retrieval-enhanced visual prompt learning for few-shot classification. *arXiv preprint arXiv:2306.02243*, 2023.

[149] Y. Zang et al. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022.

[150] S. Shen et al. Multitask vision-language prompt tuning. *arXiv preprint arXiv:2211.11720*, 2022.

[151] M. U. Khattak et al. Maple: Multi-modal prompt learning. *arXiv preprint arXiv:2210.03117*, 2022.

[152] Y. Xing et al. Class-aware visual prompt tuning for vision-language pre-trained model. *arXiv preprint arXiv:2208.08340*, 2022.

[153] O. Pantazis et al. Svl-adapter: Self-supervised adapter for vision-language pretrained models. *arXiv preprint arXiv:2210.03794*, 2022.

[154] V. Udandarao et al. Sus-x: Training-free name-only transfer of vision-language models. *arXiv preprint arXiv:2211.16198*, 2022.

[155] J. Kahana et al. Improving zero-shot models with label distribution priors. *arXiv preprint arXiv:2212.00784*, 2022.

[156] F. Peng et al. Sgva-clip: Semantic-guided visual adapting of vision-language models for few-shot image classification. *arXiv preprint arXiv:2211.16191*, 2022.

[157] R. Zhang et al. Vt-clip: Enhancing vision-language models with visual-guided texts. *arXiv preprint arXiv:2112.02399*, 2021.

[158] Z. Guo et al. Calip: Zero-shot enhancement of clip with parameter-free attention. *arXiv preprint arXiv:2209.14169*, 2022.

[159] T. Yu et al. Task residual for tuning vision-language models. In *CVPR*, pp. 10899–10909, 2023.

[160] S. Pratt et al. What does a platypus look like? generating customized prompts for zero-shot image classification. *arXiv preprint arXiv:2209.03320*, 2022.

[161] S. Menon and C. Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022.

[162] M. Wortsman et al. Robust fine-tuning of zero-shot models. In *CVPR*, pp. 7959–7971, 2022.

[163] C. Zhou et al. Extract free dense labels from clip. In *ECCV*, pp. 696–712. Springer, 2022.

[164] J. Li et al. Masked unsupervised self-training for zero-shot image classification. *arXiv preprint arXiv:2206.02967*, 2022.

[165] P. Liu et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[166] M. Jia et al. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022.

[167] N. Houlsby et al. Parameter-efficient transfer learning for nlp. In *ICML*, pp. 2790–2799. PMLR, 2019.

[168] R. Zhang et al. Pointclip: Point cloud understanding by clip. In *CVPR*, pp. 8552–8562, 2022.

[169] T. Huang et al. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *ICCV*, pp. 22157–22167, 2023.

[170] M. Xu et al. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, pp. 2945–2954, 2023.

[171] G. Chen et al. Tem-adapter: Adapting image-text pretraining for video question answer. In *ICCV*, pp. 13945–13955, 2023.

[172] T. Brown et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020.

[173] Y. Zang et al. Open-vocabulary detr with conditional matching. *arXiv preprint arXiv:2203.11876*, 2022.

[174] Z. Zhou et al. Zegclip: Towards adapting clip for zero-shot semantic segmentation. *arXiv preprint arXiv:2212.03588*, 2022.

[175] T. Lüddecke and A. Ecker. Image segmentation using text and image prompts. In *CVPR*, pp. 7086–7096, 2022.

[176] B. Li et al. Language-driven semantic segmentation. In *ICLR*, 2021.

[177] N. Zabari and Y. Hoshen. Semantic segmentation in-the-wild without seeing any segmentation examples. *arXiv preprint arXiv:2112.03185*, 2021.

[178] C. Ma et al. Open-vocabulary semantic segmentation with frozen vision-language models. *arXiv preprint arXiv:2210.15138*, 2022.

[179] F. Liang et al. Open-vocabulary semantic segmentation with mask-adapted clip. *arXiv preprint arXiv:2210.04150*, 2022.

[180] M. Xu et al. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *ECCV*, pp. 736–753. Springer, 2022.

[181] G. Ghiasi et al. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, pp. 540–557. Springer, 2022.

[182] G. Shin et al. Reco: Retrieve and co-segment for zero-shot transfer. In *NeurIPS*.

[183] J. Xie et al. Clims: Cross language image matching for weakly supervised semantic segmentation. In *CVPR*, pp. 4483–4492, 2022.

[184] Y. Lin et al. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. *arXiv preprint arXiv:2212.09506*, 2022.

[185] J. Qin et al. Freeseg: Unified, universal and open-vocabulary image segmentation. In *CVPR*, pp. 19446–19455, 2023.

[186] Z. Ma et al. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *CVPR*, pp. 14074–14083, 2022.

[187] H. A. Rasheed et al. Bridging the gap between object and image-level representations for open-vocabulary detection. In *NeurIPS*.

[188] C. Feng et al. Promptdet: Towards open-vocabulary detection using uncurated images. In *ECCV*, pp. 701–717. Springer, 2022.

[189] M. Gao et al. Open vocabulary object detection with pseudo bounding-box labels. In *ECCV*, pp. 266–282. Springer, 2022.

[190] T. Wang and N. Li. Learning to detect and segment for open vocabulary object detection. *arXiv preprint arXiv:2212.12130*, 2022.

[191] C. Lin et al. Learning object-language alignments for open-vocabulary object detection. *arXiv preprint arXiv:2211.14843*, 2022.

[192] W. Kuo et al. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022.
[193] X. Zhou et al. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, pp. 350–368. Springer, 2022.
[194] D. Huynh et al. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *CVPR*, pp. 7020–7031, 2022.
[195] M. Minderer et al. Simple open-vocabulary object detection with vision transformers. *arXiv preprint arXiv:2205.06230*, 2022.
[196] S. Zhao et al. Exploiting unlabeled data with vision and language models for object detection. In *ECCV*, pp. 159–175. Springer, 2022.
[197] Y. Long et al. P3ovd: Fine-grained visual-text prompt-driven self-training for open-vocabulary object detection. *arXiv preprint arXiv:2211.00849*, 2022.
[198] J. Xie and S. Zheng. Zsd-yolo: Zero-shot yolo detection using vision-language knowledgedistillation. *arXiv preprint arXiv:2109.12066*, 2021.
[199] D. Kim et al. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *CVPR*, pp. 11144–11154, 2023.
[200] S. Wu et al. Aligning bag of regions for open-vocabulary object detection. In *CVPR*, pp. 15254–15264, 2023.
[201] L. Wang et al. Object-aware distillation pyramid for open-vocabulary object detection. In *CVPR*, pp. 11186–11196, 2023.
[202] M. Cherti et al. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pp. 2818–2829, 2023.
[203] Y. Xian et al. Semantic projection network for zero-and few-label semantic segmentation. In *CVPR*, pp. 8256–8265, 2019.
[204] X. Zhai et al. Scaling vision transformers. In *CVPR*, pp. 12104–12113, 2022.
[205] C. Raffel et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

APPENDIX

A. STATISTICS ON VISUAL RECOGNITION VLM PUBLICATIONS

As shown in Figure 1 (in the main manuscript), we count the number of visual recognition VLM publications on Google Scholar over the recent two years. Specifically, we consider all the papers that have cited the pioneer VLM study (*i.e.*, CLIP) as potential publications and identify a publication as the visual recognition VLM study if it contains any one of keywords image classification, object detection and semantic segmentation. For the year 2023, we project the total publications based on the number of publications from 1 Jan 2023 to 30 November 2023.

B. DATASETS FOR PRE-TRAINING VLM

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], [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.

B.1. Image-Text Datasets

- **SBU [73]** contains 1M images collected from Flicker website, paired with visually relevant captions.
- **COCO Caption [74]** contains over 330k images from MS COCO [106]. It has two versions: COCO Caption c5 with 5 reference captions for 330k images and COCO Caption c40 that provides 40 reference captions for a randomly sampled subset of 5,000 images.
- **YFCC100M [75]** is a multimedia dataset containing 99.2M images and 0.8M videos with texts.
- **VG [76]** provides a multi-perspective understanding of images, *e.g.*, object-level information, scene graphs and visual question answer pairs. VG contains 108,000 images, each with 50 descriptions.
- **CC3M [77]** is an image captioning dataset which consists of about 3.3M image-text pairs from the web.
- **CC12M [79]** is introduced specifically for VLM pre-training. By relaxing the data collection pipeline used in CC3M [77], CC12M collects less precise but much larger size of data, *i.e.*, 12M image-text pairs.
- **LR [78]** is an image captioning dataset with local multi-modal annotations, where every word is localized in the image with a mouse trace segment. It contains 848,749 images with 873,107 captions.
- **WIT [80]** is a large multi-modal multilingual dataset collected from Wikipedia, which consists of 37.6M image-text pairs across 108 languages.
- **Red Caps [81]** is a image-text dataset collected from social media Reddit, which contains 12M image-text pairs covering various objects and scenes.
- **LAION400M [21]:** LAION400M consists of 400M image-text pairs filtered by CLIP [10], which also provides the data embeddings and kNN indices.
- **LAION5B [20]** contains over 5.8B image-text pairs, which consists of three parts: 2.32B English image-text pairs, 2.26B multilingual image-text pairs and 1.27B pairs without specific language.
- **WuKong [82]** is a large-scale Chinese multi-modal dataset, which contains 100M Chinese image-text pairs collected from the web.
- **CLIP [10]** is a large-scale web image-text dataset, which contains 400M image-text pairs collected form a variety of publicly available sources on the internet.
- **ALIGN [17]** is an image-text dataset, which contains 1.8B noisy image-text pairs covering board concepts.
- **FILIP [18]** is a large-scale image-text dataset with 300M image-text pairs collected from the internet.
- **WebLI [83]** is a multilingual image-text dataset collected from the web, which comprises 10B images with 12B corresponding texts across 109 languages.

B.2. Auxiliary Datasets

- **JFT3B [204]** contains nearly 3B images annotated with a noisy class hierarchy of around 30k labels.
- **C4 [205]** is a collection of about 750GB English text sourced from the public Common Crawl web scrape.
- **Object365 [85]** is a object detection dataset with 365 categories, 638K images, and $\sim$10M bounding boxes.
- **Gold-G [86]** is an object-phrase dataset for object detection, which includes 0.8M human-annotated visual grounding data curated by [86].

C. DATASETS FOR EVALUATION

Many visual recognition datasets have been adopted for VLM evaluations as shown in Table 2 (in the main manuscript) including 27 image classification datasets, 4 object detection datasets, 4 semantic segmentation datasets, 2 image-text retrieval datasets and 3 action recognition datasets. Below please find the detail of each dataset.

C.1. Datasets for Image Classification

- **Food-101** [22] is a real-world food dataset for fine-grained recognition. The dataset consists of 101,000 images, covering 101 classes. Specifically, every class contains 250 cleaned test samples and 750 purposely uncleaned training samples.
- **CIFAR-10** [23] contains a set of small images, which is commonly used for the image classification tasks. This dataset includes 60000 images with size 32 by 32, annotated with ten categories. This dataset has been divided into 5000 training samples and 1000 testing samples per class.
- **CIFAR-100** [23] is almost the same as CIFAR10, except CIFAR-100 instead contains 60000 samples with 100 categories that have been grouped into twenty super-categories.
- **Birdsnap** [100] is a fine-grained classification dataset collected from Flicker. There are 49,829 images belonging to 500 bird species, including 47,386 training images and 2433 testing images. In this dataset, every image has been labelled with a bounding box, the coordinates of 17 parts, and auxiliary attribute annotations like male, female, immature, etc.
- **SUN397** [24] is proposed for scene recognition and contains 39700 images covering 397 well-sampled categories. The scene classification performance by human is provided as the reference for the comparisons with computational methods.
- **Stanford Cars** [25] is designed for fine-grained recognition, containing 16185 images covering 196 categories. This dataset has been divided into 8,144 training samples and 8,041 testing samples.
- **FGVC Aircraft** [96] includes 10K samples spanning 100 aircraft model variants. Every samples is labeled with a tight bounding box and a hierarchical category annotation. This dataset has been equally separated into training, validation, and test subsets, where every subset contains 33 or 34 images per variant.
- **PASCAL VOC 2007 Classification** [90] is the widely-used dataset for various visual recognition tasks like detection, segmentation, and classification. There are 9963 samples covering 20 classes, including 5011 training images and 4952 testing images. Every sample in PASCAL VOC 2007 contains pixel-wise labels, object-level labels with object box and category labels.
- **Describable Textures** [99] **(DTD)** is a collection of textural images for image recognition. This dataset includes 5640 samples with forty seven categories, which has been equally separated into training, validation, and test subsets, where each subset contains 40 images per class. For each image, the main category and a list of the joint attributes are provided.
- **Oxford-IIIT PETS** [26] includes 7,349 cat and dog images with thirty seven different breeds, in which twenty five are dog breeds and twelve are cat breeds. These samples are separated into the training subset with around 1850 samples, the validation subset with about 1850 samples and testing subset with approximate 3700 samples. Every sample has been annotated with a breed annotation, a pixel-wise annotation that marks the body, and a rectangle box for locating the head.
- **Caltech-101** [89] consists of 9145 images belonging to 101 classes, Every category includes 40-80 images. For each image, the dataset provides an annotation to segment the foreground object.
- **Oxford 102 Folwers** [91] is proposed for fine-grained image classification. This dataset contains 8189 flowers images that belong to 102 species. Each category contains 40-200 samples, including the flower captured under various sizes and illumination environments. Besides, this dataset also contains pixel-wise annotations.
- **Facial Emotion Recognition 2013** [97] is collected by requesting images associated with 184 key emotional terms from Google. The dataset contains 35,887 grayscale images with a resolution of 48x48 pixels and with 7 types of emotions.
- **STL-10** [93] is a type of classification benchmark for researching on unsupervised and self-taught training. It includes 10 categories and an unsupervised training subset with 100K samples, a supervised training subset with 5K samples, and a testing subset with 8K samples.
- **EuroSAT** [104] is a set of satellite images used to benchmark the land use and land cover recognition tasks. It covers thirteen frequency bands with ten categories of 27K annotated and geo-referenced samples. Two datasets are provided including the RGB image dataset and the multi-spectral image dataset.
- **RESISC45** [101] has been proposed to benchmark Remote Sensing Image Scene Classification (RESISC). This dataset includes 31,500 sample with the image size of 256 by 256 and forty five scene categories, every category containing 700 samples. Besides, RESISC45 covers a wide range of spatial resolutions from 20cm to over 30m per pixel.
- **GTSRB** [94] is a dataset for traffic signs classification, containing 50,000 images taken from various street scenes in Germany. It is classified into 43 categories, including a training subset with 39,209 samples and a testing subset with 12,630 samples.
- **Country211** [10] is an image classification dataset for geolocation evaluation, which is a subset of the YFCC100M dataset. For each country, there are one hundred and fifty train samples, fifty validation samples, and one hundred test samples.
- **PatchCamelyon** [103] includes 327,680 RGB images with the size of 96 by 96 from Camelyon16, with a training subset with 75% samples, a validation subset with 12.5% samples, and a testing subset with 12.5% samples. Every sample has been labelled with a binary annotation showing if it contains the metastatic

tissue.

- **Hateful Memes** [27] has been proposed for hateful meme classification (*i.e.*, image with text) created by Facebook AI. It includes over 10k memes annotated with either the hateful label or the non-hateful label.
- **Rendered SST2** [10] has been proposed for benchmarking optical character recognition. It includes 2 categories (the categoryies of positives and negatives). This dataset has been separated into 3 subsets: a training subset with 6920 samples, a validation subset with 872 samples, and a test subset with 1821 samples.
- **ImageNet-1k** [40] includes about 1.2M samples that are uniformly distributed across the one thousand categories. The category annotation of ImageNet-1k follows WordNet hierarchy and every sample is annotated with one category label. Bisides, ImageNet-1k is one of the most popular image classification benchmarks.
- **CLEVR Counts** [102] is a subset of CLEVR dataset, which is designed for visual question answering to evaluate the ability to perform visual reasoning. The counting tasks include 2000 training samples and 500 test samples.
- **SVHN** [92] is a dataset for recognizing digits and numbers in real-world images which are collected from Google Street View images. It consists of about 600,000 images and all digits are cropped from the images and resized to a fixed resolution of 32x32 pixels.
- **IIIT5k** [95] contains 5,000 cropped word images collected from Google image search by using search keyword such as signboards, house name plates, and movie posters, etc. The dataset is split into two parts, *i.e.*, 2,000 word images for training and 3,000 word images for validation, respectively.
- **Rendered SST2** [98] is sentiment classification dataset which consists of two sentiment categories, *i.e.*, negative and positive. The sentences in the dataset are extracted from Stanford Sentiment Treebank dataset.

C.2. Datasets for Action Recognition

- **UCF101** [29] has been proposed for benchmarking human action recognition with videos. It includes about 13k video clips of 101 actions, which are collected from YouTube. The video clips in the dataset have a resolution of 320x240 pixels and a frame rate of 25 FPS.
- **Kinetics700** [30] is a video dataset for recognizing human action. It consists of about 65,000 video clips with 700 human actions, where each action category has more than 700 video clips lasting around 10 seconds.
- **RareAct** [28] is a video dataset designed for identifying rare actions such as "Unplug Oven" and "fry phone". This dataset aims to evaluate action recognition models on unlikely combinations of common action verbs and object nouns. It contains 122 human actions, where the verbs and object nouns in actions are rarely co-occurring together in HowTo100M.

C.3. Datasets for Semantic Segmentation

- **ADE20k** [111] is a semantic segmentation dataset which consists of 150 classes. It consists of a training subset with 25,574 samples and a validation subset with 2,000 samples.
- **PASCAL VOC 2012 Segmentation** [90] contains 20 categories including vehicles, household and animals. This dataset includes a training subset with 1,464 samples and a testing subset with 1,449 samples, all of which are with pixel-wise annotations.
- **PASCAL Content [109]:** PASCAL Content is an extension of PASCAL VOC 2010 detection dataset [90], which contains more than 400 categories with pixel-wise annotations. It has 4,998 training images and 1,449 validation images.
- **Cityscapes [110]:** Cityscapes is a dataset for the visual recognition on street scenes. This dataset includes a training subset with 2,975 samples and a testing subset with 500 samples, all of which are with pixel-wise annotations of 19 categories.

C.4. Datasets for Object Detection

- **MS COCO [106]:** MS COCO Dataset is a dataset for object detection. It consists of two versions: MS COCO 2014 contains 83,000 training images and 41,000 validation images with bounding box annotations of 80 categories and MS COCO 2017 contains 118,000 training images and 5,000 validation images with bounding box annotations of 80 categories.
- **ODinW [108]:** ODinW is a benchmark to evaluate the task-level transfer ability of pre-trained vision models, which consist of 35 free public Object Detection datasets in various domains. The dataset contains 132k training images and 20K testing images belonging to 314 concepts. Also, many of the 35 tasks have very limited (less than 100) training images, which makes it extremely difficult for standard detectors without any pre-training.
- **LVIS [107]:** LVIS is a large vocabulary dataset for long-tailed instance detection/ segmentation. The dataset contains 1203 categories with federated human annotations on 100K images.

C.5. Datasets for Image and Text Retrieval

- **Flickr30k [105]:** Flickr30K is a dataset for automatic image description and grounded language understanding. It contains 31,000 images collected from Flickr, where each image is provided with 5 captions.
- **COCO Caption [74]:** COCO Caption contains over 330k images from MS COCO [106]. It has two versions: COCO Caption c5 with 5 reference captions for 330k images and COCO Caption c40 that provides 40 reference captions for a randomly sampled subset of 5,000 images.

Jingyi Zhang received her B.Sc. degree in electronic information science and technology from the University of Electronic Science and Technology of China (UESTC) and M.Sc. degree in signal processing from the Nanyang Technological University (NTU). She is currently a Research Associate and Ph.D. student with School of Computer Science and Engineering, NTU. Her research interests include computer vision, object detection.

Jiaxing Huang received his B.Eng. and M.Sc. in EEE from the University of Glasgow, UK, and the Nanyang Technological University (NTU), Singapore, respectively. He is currently a Research Associate and Ph.D. student with School of Computer Science and Engineering, NTU, Singapore. His research interests include computer vision and machine learning.

Sheng Jin is currently a Research Fellow at Nanyang Technology University (NTU), Singapore. Before that, he received his B.Sc. degree in Applied Mathematics from Harbin Institute of Technology and the Ph.D. degree in Computer Science and Technology om Harbin Institute of Technology. His research interests include computer vision and machine learning.

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