Large-scale Multi-Modal Pre-trained Models: A Comprehensive Survey

Xiao Wang[1,2], Guangyao Chen[1,3], Guangwu Qian[1], Pengcheng Gao[1], Xiao-Yong Wei[1,4], Yaowei Wang(✉)[1], Yonghong Tian(✉)[1,3] and Wen Gao[1,3]

[1]Pengcheng Laboratory, Shenzhen 518055, China.
[2]School of Computer Science and Technology, Anhui University, Hefei 230601, China.
[3]School of Computer Science, Peking University, Beijing 100871, China.
[4]College of Computer Science, Sichuan University, Chengdu 610065, China.

Abstract

With the urgent demand for generalized deep models, many pre-trained big models are proposed, such as BERT, ViT, GPT, etc. Inspired by the success of these models in single domains (like computer vision and natural language processing), the multi-modal pre-trained big models have also drawn more and more attention in recent years. In this work, we give a comprehensive survey of these models and hope this paper could provide new insights and helps fresh researchers to track the most cutting-edge works. Specifically, we firstly introduce the background of multi-modal pre-training by reviewing the conventional deep learning, pre-training works in natural language process, computer vision, and speech. Then, we introduce the task definition, key challenges, and advantages of multi-modal pre-training models (MM-PTMs), and discuss the MM-PTMs with a focus on data, objectives, network architectures, and knowledge enhanced pre-training. After that, we introduce the downstream tasks used for the validation of large-scale MM-PTMs, including generative, classification, and regression tasks. We also give visualization and analysis of the model parameters and results on representative downstream tasks. Finally, we point out possible research directions for this topic that may benefit future works. In addition, we maintain a continuously updated paper list for large-scale pre-trained multi-modal big models: https://github.com/wangxiao5791509/MultiModal_BigModels_Survey.

Keywords: Multi-modal, Pre-trained Model, Information Fusion, Representation Learning, Deep Learning

1 Introduction

Along with the breakthroughs of recognition performance of AlexNet [1] on the ImageNet competition [2], the artificial intelligence have developed greatly. Many representative deep neural networks are proposed, such as VGG [3], ResNet [4], Inception [5], LSTM [6]. The researchers usually collect and annotate some samples for their task, and train their models based on pre-trained backbones on large-scale datasets (such as ImageNet [2] for computer vision, Glove [7] and Skip-thought vectors [8] for natural language processing). Many tasks can be solved well in such an end-to-end manner compared with traditional handcrafted features, such as object detection, segmentation, and recognition. However, the generalization ability of obtained deep model is still limited. Collecting and annotating a larger dataset can address these issues to some extent, but this procedure is expensive and tedious.

To address this issue, Ashish et al. propose the Transformer network [9] which achieves new SOTA (State-Of-The-Art) performance on machine translation task. After that, the self-supervised pre-training on large-scale corpus, then, fine-tuning on downstream tasks attracts more and more researchers' attention. Many pre-trained big models are proposed by following such paradigm, such as BERT [10], GPT [11, 12], T5 [13], XLNet [14] which also trigger new research highlights of pre-training in CV community. More and more large-scale NLP and CV models demonstrate the powerful effect by pretrain-and-finetuning paradigm, including ViT [15] and Swin-Transformer [16].

Although the progress brings new impetus to the development of artificial intelligence, however, the issues caused by the defect of single modality are still hard to solve. Researchers attempt to incorporate more modalities to bridge the data gap for deep models. Many multi-modality fusion based tasks are also explored in a traditional deep learning manner, such as RGB, Depth, Natural Language, Point Cloud, Audio, Event stream, etc. Many large-scale pre-trained multi-modal models [17–23] are proposed which set new SOTA on downstream tasks one after another, as shown in Fig. 1. In this paper, we give a comprehensive review of these works which target to help the new researchers who are interested in this area to understand the history and latest developments quickly.

Organization of our review. In this paper, we firstly review the background of multi-modal pre-training technique in Section 2, from the traditional deep learning paradigm to pre-training in single modality tasks, including natural language processing, computer vision, and automatic speech processing. Then, we focus on MM-PTMs and describe the task definition, key challenges, and benefits, in Section 3.1 and 3.2. The key components are also reviewed in the following sub-sections, including large-scale data, network architectures, optimization objectives, and knowledge-enhanced pre-training. To validate the effectiveness of pre-trained models, many downstream tasks are used for quantitative assessment. In Section 4, we provide detailed reviews on the task definition and evaluation metrics of these tasks. In Section 5, we review the model parameters and hardware for training and also report the experimental results of several representative downstream tasks. Finally, in Section 6, we conclude this survey and propose multiple research directions needed to be studied. The architecture of this survey is visualized in Fig. 2.

Difference from existing reviews. Although there are already two surveys [24, 25] proposed for MM-PTMs, the difference between our survey and existing ones can be summarized as follows:

- **Scope:** Existing multi-modal surveys [24, 25] focus on vision-language only, however, the multi-modal information problem is a wider research topic. This paper is more comprehensive than the aforementioned reviews by introducing more modalities, such as audio, video, table, etc.
- **Timeliness:** This paper introduces the latest datasets and algorithms (from the year 2019 to June 2022) proposed for multi-modal pre-training which is a long survey, meanwhile, their work belongs to short paper.
- **New insights to MM-PTMs:** By classifying and analyzing the existing MM-PTMs from different perspectives, this article can help readers master the cutting-edge methods and techniques from both detailed and high-level perspectives. In addition, our proposed research directions on the MM-PTMs are deliberate and will provide new clues for the follow-up research.

2 Background

2.1 Conventional Deep Learning

With the release of AlexNet [1], a series of deep learning models are proposed in the artificial intelligence community. These deep models show better capabilities for fitting complex data than conventional machine learning models. From the perspective of its development (LeNet [51] → AlexNet [1] → VGG [3] → ResNet [4] → DenseNet [52]), we can find that their architectures become deeper and deeper, and the corresponding performance accordingly becomes better. The success of these approaches is supported by large-scale annotated training data, such as the ImageNet [2] for the classification task. The scale of used data is much larger than traditional methods, but it's still limited. The pursuit

Table 1 Summary of related single- and multi-modal pre-training surveys. SC and DC denotes Single Column and Double Column. Pub. is short for Publication.

No.	Title	Year	Pub.	Topic	Pages
01	A short survey of pre-trained language models for conversational ai-a new age in nlp [26]	2020	ACSWM	NLP	DC, 4
02	A Survey of Controllable Text Generation using Transformer-based Pre-trained Language Models [27]	2022	arXiv	NLP	SC, 34
03	A Survey of Knowledge Enhanced Pre-trained Models [28]	2021	arXiv	KE	DC, 20
04	A Survey of Knowledge-Intensive NLP with Pre-Trained Language Models [29]	2022	arXiv	KE	DC, 8
05	Commonsense Knowledge Reasoning and Generation with Pre-trained Language Models: A Survey [30]	2022	arXiv	KE	DC, 11
06	A survey on contextual embeddings [31]	2020	arXiv	NLP	DC, 13
07	Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing [32]	2021	arXiv	NLP	SC, 46
08	Pre-trained Language Models in Biomedical Domain: A Systematic Survey [33]	2021	arXiv	NLP	SC, 46
09	Pre-trained models for natural language processing: A survey [34]	2020	SCTS	NLP	DC, 26
10	Pre-Trained Models: Past, Present and Future [35]	2021	AI Open	NLP, CV, MM	DC, 45
11	Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey [35]	2021	arXiv	NLP	DC, 49
12	A Survey of Vision-Language Pre-Trained Models [36]	2022	arXiv	MM	DC, 9
13	Survey: Transformer based video-language pre-training [37]	2022	AI Open	CV	DC, 13
14	Vision-Language Intelligence: Tasks, Representation Learning, and Large Models [38]	2022	arXiv	MM	DC, 19
15	A survey on vision transformer [39]	2022	TPAMI	CV	DC, 23
16	Transformers in vision: A survey [40]	2021	CSUR	CV	SC, 38
17	A Survey of Visual Transformers [41]	2021	arXiv	CV	DC, 21
18	Video Transformers: A Survey [42]	2022	arXiv	CV	DC, 24
19	Threats to Pre-trained Language Models: Survey and Taxonomy [43]	2022	arXiv	NLP	DC, 8
20	A survey on bias in deep NLP [44]	2021	AS	NLP	SC, 26
21	A Survey of Controllable Text Generation using Transformer-based Pre-trained Language Models [27]	2022	arXiv	NLP	SC, 34
22	An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-Trained Language Models [45]	2021	arXiv	NLP	DC, 21
23	A multi-layer bidirectional transformer encoder for pre-trained word embedding: A survey of BERT [46]	2020	CCDSE	NLP	DC, 5
24	Survey of Pre-trained Models for Natural Language Processing [47]	2021	ICEIB	NLP	DC, 4
25	A Roadmap for Big Model [48]	2022	arXiv	NLP, CV, MM	SC, 200
26	Vision-and-Language Pretrained Models: A Survey [49]	2022	IJCAI	MM	DC, 8
27	Multimodal Learning with Transformers: A Survey [50]	2022	arXiv	MM	DC, 23

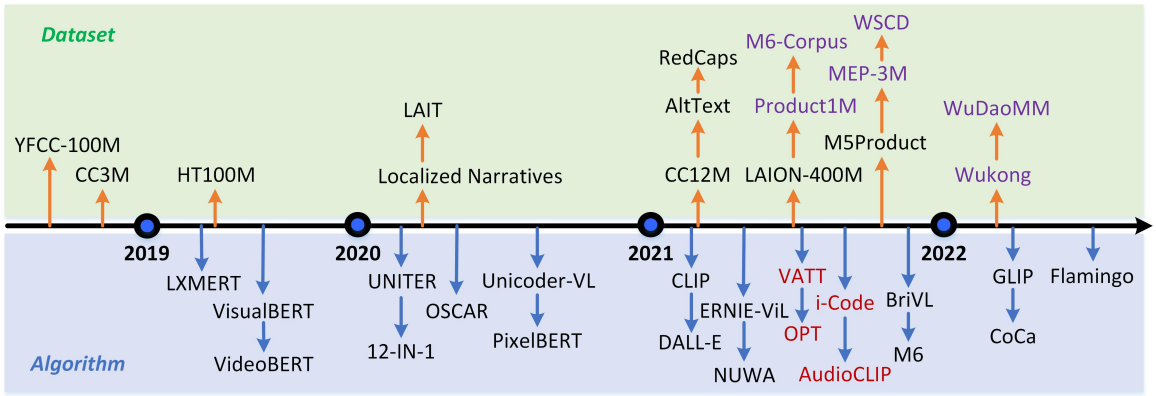


Fig. 1 The chronological milestones on multi-modal pre-trained big models from 2019 to the present (June 2022), including multi-modal datasets (as shown by the orange arrow) and representative models (as shown by the blue arrow). The purple font indicates that the dataset contains Chinese text (other datasets contain English text). The models highlighted in wine red are trained on more than two modalities.

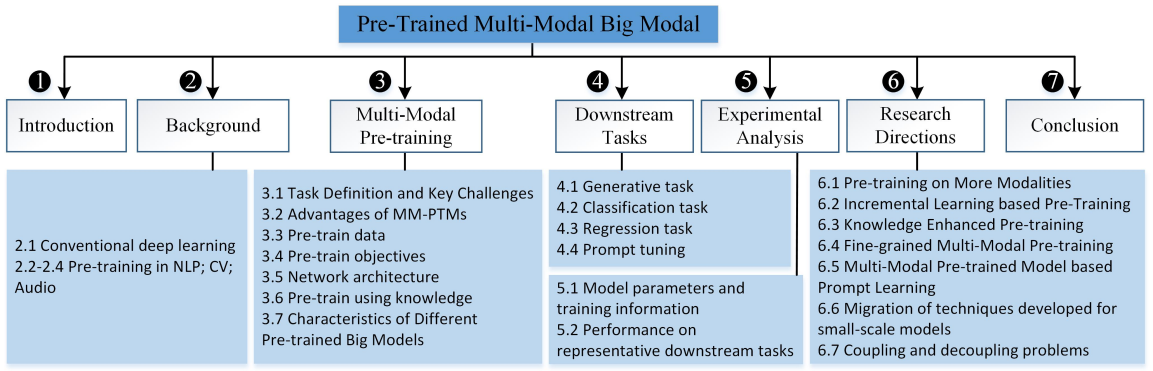


Fig. 2 The overall framework of this survey.

of robustness and generalization performance of machine learning models has never stopped.

Recently, the results of large-scale pre-trained models obtained by pre-training on massive data are constantly refreshing people's cognition of artificial intelligence. Compared with previous small-scale deep learning methods, pre-trained big models show obvious advantages in Natural Language Processing (NLP), Computer Vision (CV), and Multi-Modal fields. Such a pre-training scheme take full advantage of the large-scale unlabeled data, therefore, getting rid of expensive annotation costs. Therefore, the study of large-scale pre-trained models is a feasible and necessary way to explore real intelligence.

2.2 Pre-training in Natural Language Processing

The large-scale pre-trained models [29, 43, 44, 53–56] first appeared in the NLP field. Their success is mainly attributed to self-supervised learning and network structures like Transformer [9]. Specifically, the advent of Bidirectional Encoder Representations (BERT) [10] based on self-supervised learning has led to revolutionary performance improvements on a wide variety of downstream tasks by fine-tuned on fewer training data [57]. Generative Pre-trained Transformers (GPT) [12, 58, 59] further extends the number of parameters and the training data for better performance. Note that, the GPT-3 [12] has ten times more parameters than TuringNLP [60]. It can not only better fulfill the functions of general NLP tasks, but also has some mathematical calculation ability. The success of the GPT-3 model has made

it widely used in various fields, such as search engines, chatbots, music composition, graphics, and coding. XLNet [14] is developed based on a generalized permutation language modeling objective, which achieves unsupervised language representation learning. PanGu-α [61] is a large-scale pre-trained Chinese model with 200 billion parameters and implemented based on MindSpore Auto-parallel. NEZHA [62] is another Chinese pre-trained big model based on BERT proposed by Wei et al. More large-scale pre-trained models for NLP can be found in surveys [27, 34].

2.3 Pre-training in Computer Vision

Inspired by the revolutionary advancement of Transformer for NLP tasks, many large-scale Transformer-based vision models are also proposed in recent years. Chen et al. [63] attempt to auto-regressively predict pixels using a sequence Transformer. The model obtained by pre-training on the low-resolution ImageNet dataset demonstrates strong image representations. The ViT (Vision Transformer) model [64] directly adopts the pure Transformer to handle the sequence of image patches for classification. Many new SOTA performances are achieved on several downstream CV tasks, including object detection [65], semantic segmentation [66], image processing [67], video understanding [67]. The Swin-Transformer [16] is another milestone for computer vision, as a hierarchical Transformer, it adopts shifted windows for representation learning.

For the pre-training methods, the Masked Image Modeling (MIM) [63, 64] is proposed to learn rich visual representations via masked parts prediction by conditioning on visible context. MIM provides another direction for the exploration of the visual large-scale pre-training model. He et al. propose the MAE [68] to re-explore pixel regression in MIM and show more comparable performance on multiple image recognition tasks. BEiT [69] greatly improves MIM's performance via masked visual token prediction, and PeCo [70] finds injecting perceptual similarity during visual codebook learning benefits MIM pre-trained representation.

2.4 Pre-training in Audio and Speech

As one of the most popular modalities, the audio and speech based pre-training also draws the researcher's attention. For example, the wav2vec [71] is the first work that applies contrastive learning to improve supervised speech recognition by learning the future raw audio based on the past raw audio. The vq-wav2vec [71] uses context prediction tasks from wav2vec to learn the representations of audio segments. Discrete-BERT [72] is BERT-style model by finetuning the pre-trained BERT models on transcribed speech. HuBERT [73] uses self-supervised speech learning where an offline clustering step is used to generate discrete labels of masked speech signals. wav2vec 2.0 [74] solves a contrastive task to predict the masked latent representation. w2v-BERT [75] uses contrastive learning and masked speech modeling simultaneously, where a model predicts discretized speech tokens and another model solves a masked prediction task.

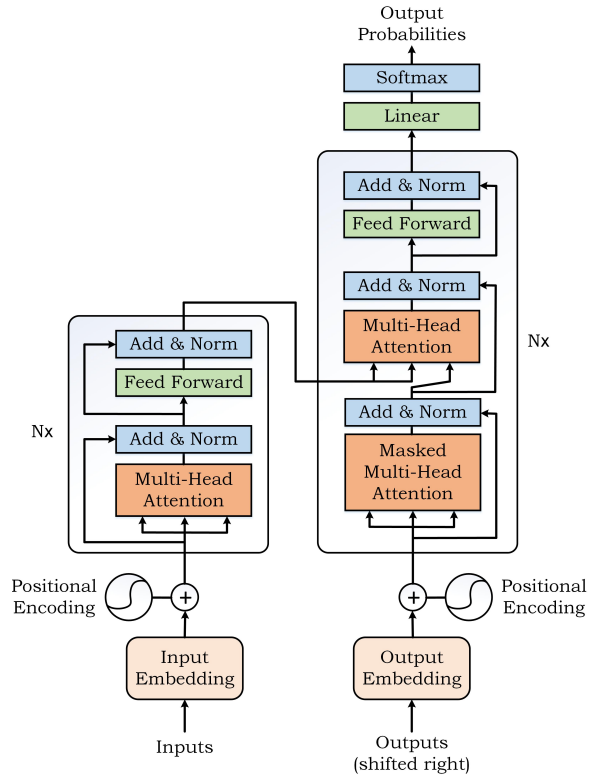


Fig. 3 The detailed network architecture of Transformer network [9].

3 Multi-Modal Pre-training

3.1 Task Definition and Key Challenges

Task Definition. Usually, the deep neural networks are trained on a large-scale dataset, for example, the widely used residual network [4] are pre-trained using a classification task on the ImageNet dataset [2]. In contrast, the multi-modal pre-training big models are usually trained on a massive training dataset. Usually, these data are not annotated with labels due to the scale are too large to annotate. On the other hand, the parameters need to reach a certain scale. As illustrated in Fig. 4, the multi-modal data, big model, and computing power are tightly connected. All in all, with the support of computing power, the multi-modal pre-training usually denotes the task that the multi-modality model with huge parameters pre-trained on the massive multi-modal data in an unsupervised way.

Key Challenges. It is challenging to attain a great multi-modal pre-training big model according to aforementioned process. More in detail, we summarize the following key challenging factors:

• **Acquisition and clean of large-scale multi-modal data.** The multi-modal data is one of the most important elements in MM-PTMs. The collection of multi-modal data is significantly harder than the single one, due to the scarce of multi-modal imaging devices. The frequently used multi-modal cameras are usually covers two modalities only, such as RGB-Depth, RGB-Thermal, RGB-Radar, RGB-Event cameras, etc. Most of current MM-PTMs are vision-language models, because of the easy access to image and text data from the Internet. But the additional cleaning of these data is also necessary due to the noisy samples.

• **Design of network architectures for large-scale multi-modal pre-training.** The network architecture is another key component for multi-modal pre-training. The networks used for feature encoding of multiple input modalities are worthy carefully tailored, as different modalities may have their own features and particular networks are needed. For example, the Transformer or CNN are suggested for image and text modality, the spiking networks can be used for event streams. Another problem is the design of multi-modal fusion or cross-modality matching modules. Whether similar modules designed for small-scale multi-modal tasks work for large-scale pre-trained models or not are still remain to be verified.

• **Design of pre-training objectives.** Due to the massive unlabelled multi-modal data, the pre-training tasks usually need to be done in an unsupervised learning manner. Many current works adopt the masked region prediction for each modality as their learning objective. Obviously, the objectives for multi-modal tasks can be directly borrowed from single-modality pre-training, however, the pre-training objectives designed for the multi-modal tasks are also necessary, intuitive and effective. The widely used contrastive learning, modality based matching, and modality translation are all valid and meaningful attempts. How to design new multi-modal pre-training objectives is one of the most challenging tasks for MM-PTMs.

• **Support of large-scale computing power.** The training for traditional deep neural networks can be executed on a server with limited number of GPUs. In contrast, the MM-PTMs needs more computing power due to the large-scale multi-modal data and the super large-scale model parameters. Therefore, the first thing is to prepare a supercomputing device and the subsequent model training also requires a lot of power to support.

• **Skills on parameter tuning.** It is never a simple task to train an effective large model considering aforementioned challenging factors. The tricks used for training the neural networks are also very important. As the research and techniques for the small scale pre-training are relatively more mature, however, there is less accumulation of experience on large-scale pre-training techniques.

3.2 Advantages of MM-PTMs

Compared with *single modality pre-trained big models*, the MM-PTMs are more suitable for practical application scenarios. Specifically, the problems like multi-modal collaborative generation, modal completion, cross-domain retrieval, etc, can be addressed well using MM-PTMs. Also, the multi-modal data contains more information which can make up for the defects of a

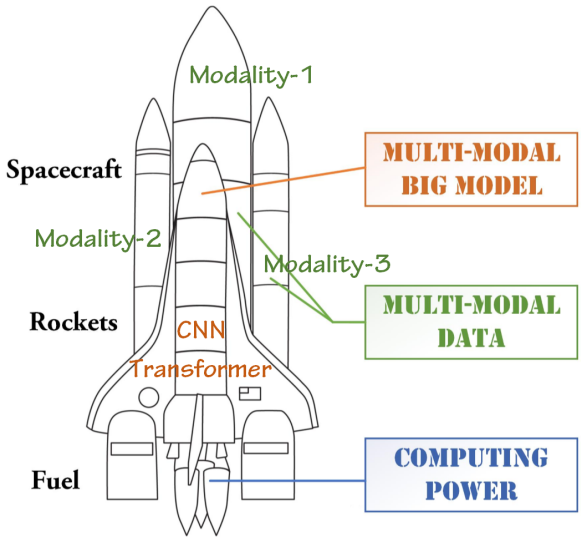


Fig. 4 The relations between multi-modal data, model, and computing power.

single modality. Therefore, the MM-PTMs can help extracting the common features of multi-modalities. Many recent works demonstrate that the utilization of MM-PTMs indeed brings in the additional prior knowledge [76–78].

Compared with *small-scale multi-modal models*, the generalizability of MM-PTMs which are obtained by self-supervised/unsupervised learning can be improved significantly. As some prior knowledge is only contained in massive big data, and a small amount of artificially selected annotated data is biased, therefore, it is hard for the small-scale models to master such knowledge.

3.3 Pre-training Data

As shown in Table 2, many large-scale multi-modal datasets are proposed for the pre-training task. In this subsection, we will briefly introduce these datasets to help readers quickly master the data information for pre-training.

• **SBU Captions** [79] is originally collected by querying Flickr [1] using plentiful query terms. Then, they filter the obtained large-scale but noisy samples to get the dataset, which contains more than 1M images with high-quality captions.

• **Flickr30k** [80] is obtained by extending Hodosh et al. [110] 's corpus with 31,783 photographs collected from Flickr. These images cover everyday activities, events, and scenes. Five sentences are annotated for each collected image via crowdsourcing, therefore, Flickr30k contains 158,915 captions.

• **COCO** [111] is developed based on MS-COCO dataset [111] which contains 123,000 images. The authors recruit the Amazon Mechanical Turk [2] to annotate each image with five sentences.

• **Visual Genome** [82] is proposed to help develop machine learning models that can understand the image by mining the interactions and relationships between objects. Therefore, they perform well on the cognitive tasks, such as the image description and visual question answering, etc. Statistically, the Visual Genome dataset contains more than 108K images and each image has about 35 objects, 26 attributes, 21 pairwise relationships.

• **VQA v2.0** [83] is proposed to reduce the language biases that existed in previous VQA datasets which contain about 1.1M image-question samples and 13M associated answers on 200K visual images from the COCO dataset.

• **FashionGen** [84] contains 325,536 high-resolution images (1360 × 1360), each image has a paragraph-length descriptive captions sourced from experts. Six different angles are photographed for all fashion items.

• **CC3M** [85] is a dataset annotated with conceptual captions proposed in 2018. The image-text samples are mainly collected from the web, then, about 3.3M image-description pairs remained after some necessary operations, such as extract, filter, and transform.

• **CC12M** [88] is the outcome of urgent need of MM-PTMs for large-scale data. The released CC3M dataset is far failed to meet the demand, therefore, the authors further relax the filters used in CC3M for the image and text cleaning. Correspondingly, a four times larger dataset CC12M can be obtained with a slight loss of accuracy.

• **GQA** [86] is mainly proposed for visual reasoning and compositional question answering. A robust question engine is carefully refined by considering *content* and *structure* information. Then, the associated semantic representations are adopted to greatly reduce biases within the

[1] https://www.flickr.com/

[2] https://www.mturk.com/

Table 2 An overview of multi-modal datasets proposed for large-scale pre-training. Lang. and Ava. is short for Language and Available, respectively.

No.	Datasets	Year	Scale	Modal	Lang.	Ava.	URL
01	SBU Captions [79]	2011	1M	image-text	English	✓	Link
02	Flickr30k [80]	2014	145K	image-text	English	✓	Link
03	COCO [81]	2014	567K	image-text	English	✓	Link
04	Visual Genome [82]	2017	5.4M	image-text	English	✓	Link
05	VQA v2.0 [83]	2017	1.1M	image-text	English	✓	Link
06	FashionGen [84]	2018	300k	image-text	English	✓	Link
07	CC3M [85]	2018	3M	image-text	English	✓	Link
08	GQA [86]	2019	1M	image-text	English	✓	Link
09	LAIT [87]	2020	10M	image-text	English	×	-
10	CC12M [88]	2021	12M	image-text	English	✓	Link
11	AltText [89]	2021	1.8B	image-text	English	×	-
12	TVQA [90]	2018	21,793	video-text	English	✓	Link
13	HT100M [91]	2019	136M	video-text	English	✓	Link
14	WebVid2M [92]	2021	2.5M	video-text	English	✓	Link
15	YFCC-100M [93]	2015	100M	image-text	English	✓	Link
16	LAION-400M [94]	2021	400M	image-text	English	✓	Link
17	RedCaps [95]	2021	12M	image-text	English	✓	Link
18	Wukong [96]	2022	100M	image-text	Chinese	✓	Link
19	CxC [97]	2021	24K	image-text	English	✓	Link
20	Product1M [98]	2021	1M	image-text	Chinese	✓	Link
21	WIT [99]	2021	37.5M	image-text	Multi-lingual	✓	Link
22	JFT-300M [100]	2017	30M	image-text	English	×	-
23	JFT-3B [101]	2021	3000M	image-text	English	×	-
24	IG-3.5B-17k [102]	2018	350M	image-text	English	×	-
25	M6-Corpus [103]	2021	60M	image, image-text	Chinese	×	-
26	M5Product [104]	2021	6M	image, text, table video, audio	English	✓	Link
27	Localized Narratives [105]	2020	849k	image, audio, text, mouse trace	English	✓	Link
28	RUC-CAS-WenLan [106]	2021	30M	image-text	Chinese	×	-
29	WuDaoMM [107]	2022	600M	image-text	Chinese	✓	Link
30	MEP-3M [108]	2021	3M	image-text	Chinese	✓	Link
31	WSCD [109]	2021	650M	image-text	Chinese	×	-

dataset and control for its question type composition. Finally, a balanced dataset with 1.7M samples is obtained.

• **LAIT** [87] (Large-scale weAk-supervised Image-Text) is a large-scale image-text dataset collected from the Internet in a weak-supervised manner. It contains about 10M visual images, and each image has a corresponding natural language description which contains about 13 words.

• **AltText** [89] is collected by following the rules for constructing Conceptual Captions dataset [85]. To get a large-scale dataset (1.8B image-text pairs), the authors only apply minimal frequency-based filtering for data cleaning. Although the obtained resulting dataset is noisy, the big models obtained by pre-training on this dataset still beats many SOTA works on many downstream tasks.

• **TVQA** [90] is build based on six long-running TV shows from 3 genres, including sitcoms, medical dramas, and crime drama. Then, the Amazon Mechanical Turk is used for VQA collection of video clips. Finally, this dataset contains about 152,545 question-answer pairs from 21,793 video clips.

• **HT100M** [91] contains about 136 million video clips, which are collected from 1.22 million narrated instructional videos. The content of these videos are mainly focus on humans with a total of 23,000 various tasks. The language description for each clip is an automatically transcribed narration. Therefore, the video and text are weakly-paired, compared with other captioning datasets.

• **WebVid2M** [92] is a video-text captioning dataset which contains over two million video alt-text pairs. These data are collected from the Internet following a similar procedure to CC3M

dataset. The authors find that more than 10% of CC3M images are thumbnails from videos, therefore, they scrape these video sources (a total of 2.5M text-video pairs) and create the WebVid2M dataset.

• **YFCC-100M** [93] totally contains 100 million media objects (99.2 million photos, 0.8 million videos) collected from Flickr, the time span of these videos from 2004 and 2014. Note that the YFCC100M dataset is constantly evolving, various expansion packs are unscheduled released.

• **LAION-400M** [94] contains 400 million image-text pairs which is released for vision-language related pre-training. It is worthy to note that this dataset is filtered using CLIP [77] which is a very popular pre-trained vision-language model.

• **RedCaps** [95] is a large-scale dataset with 12M image-text samples collected from 350 subreddits. The authors firstly define the range of subreddit, then, filter the image post and clean the captions. The ethical issue is also considered when building the dataset, and the problematic images are filtered according to privacy, harmful stereotypes, etc.

• **Wukong** [96] is the currently largest dataset collected from the Internet which contains 100 million image-text pairs. A list of 200K queries is maintained to ensure the collected samples cover diverse visual concepts. These queries are fed into the Baidu Image Search Engine, then, the image and its corresponding captions can be obtained. Note that each query can get at most 1000 samples to keep a balance between different queries and a series of filtering strategies are adopted for the final Wukong dataset.

• **CxC** [97] is extended based on MS-COCO dataset by rating existing and new pairs with continuous (0-5) semantic similarity. In general, the CxC contains human ratings for 267,095 pairs which is a significant extension in scale and detail. It can be used for a variety of tasks, such as the image-text, text-text, and image-image retrieval, etc.

• **Product1M** [98] contains 1,182,083 image-caption pairs, 458 categories, 92,200 instance. Each image contains about 2.83 objects. Different from regular object detection benchmark datasets, this dataset obtains the instance locations in a paste manner. They first segment the target object, then, paste them into other images based on a given bounding box. It can be used for multiple tasks, including weak-supervised, multi-modal, and instance-level retrieval.

• **WIT** [99] is constructed by crawling on Wikipedia [3]. Then, a set of rigorous filtering operations are executed on these data which finally resulting the dataset containing over 37.5 million image-text sets. Note that, the WIT dataset contains multi-lingual, in contrast, other image-text datasets only contain single lingual (for example, English or Chinese).

• **JFT-300M** [100] contains about 300M images and 375M labels, and each image has about 1.26 labels. Note that, 18291 categories are annotated in this dataset, including 1165 animals and 5720 vehicles, etc. A rich hierarchy is formed according to these categories. It is worthy to note that this dataset is not available online.

• **JFT-3B** [101] is also an internal Google dataset, which contains about 3 billion images. These samples are annotated in a semi-automatic way with a class hierarchy of 30,000 labels. In other words, this dataset contains large amount of noisy samples. Note that, this dataset is also not available online.

• **IG-3.5B-17k** [102] is constructed for weakly supervised pre-training by collecting images from Instagram [4]. Similar with JFT-300M [100] and JFT-3B [101], the dataset is also inaccessible and can only be used within the Facebook.

• **M6-Corpus** [103] is specifically constructed for the pre-training of vision-Chinese big model M6 [103]. The samples are collected from various sources, such as the product description, community question answering, forum, etc. It contains 60.5M images and 111.8B tokens.

• **M5Product** [104] is a benchmark dataset specifically proposed for E-commerce. It contains 6 million multi-modal samples which cover 6,000 categories, 5,000 attributes, and five modalities, including the visual image, table, video, language description, and audio. It is worthy to note that the M5Product dataset is different from standard multimodal datasets which have completely paired samples, that is to say, each sample may only contain only a subset of modalities. It also has a challenging long-tailed distribution issue.

[3] https://www.wikipedia.org/
[4] https://www.instagram.com/

• **Localized Narratives** [105] is proposed by Jordi et al. in 2020, which provides a new form of multi-modal image annotations for the connection of vision and language. The image and corresponding spoken description, textual description, and mouse trace are all embodied in this dataset which provides dense grounding between language and vision. It contains 849k images and covers the whole COCO, Flickr30k, and ADE20K [112] datasets and 671k images of Open Images.

• **RUC-CAS-WenLan** [106] is obtained by crawling multi-source image-text data and totally contains about 30M image-text pairs. These samples covers a wide range of topics and categories, such as the sports, entertainment, news, art, and culture, etc. It plays a fundamental role in the WenLan project and supports the training of the BriVL model [106].

• **WSCD** [109] (Weak Semantic Correlation Dataset) is a multi-source dataset, which contains large-scale image-text data samples (650 million). The English texts are all translated into Chinese to support the pre-training of BriVL.

• **MEP-3M** [108] is a large-scale image-text dataset collected from several Chinese large E-commerce platforms which contains 3 million image-text pairs of products and 599 classes. Another key feature of this dataset is the hierarchical category classification, in detail, it covers 14 classes, 599 sub-classes, and 13 sub-classes have further sub-subclasses.

3.4 Pre-training Objectives

How to design the learning objectives is a very important step for multi-modal pre-training. Currently, the following learning objectives are proposed, including contrastive loss, generative loss, etc.

• **Contrastive loss (CS)** function usually constructs positive and negative training samples which is widely used in dual-modality. For example, CLIP [77], ALIGN [21] are all trained using contrastive learning loss. The authors of VinVL [113] adopt the *3-way contrastive loss* for the pre-training to replace the binary contrastive loss function utilized in the Oscar model [17].

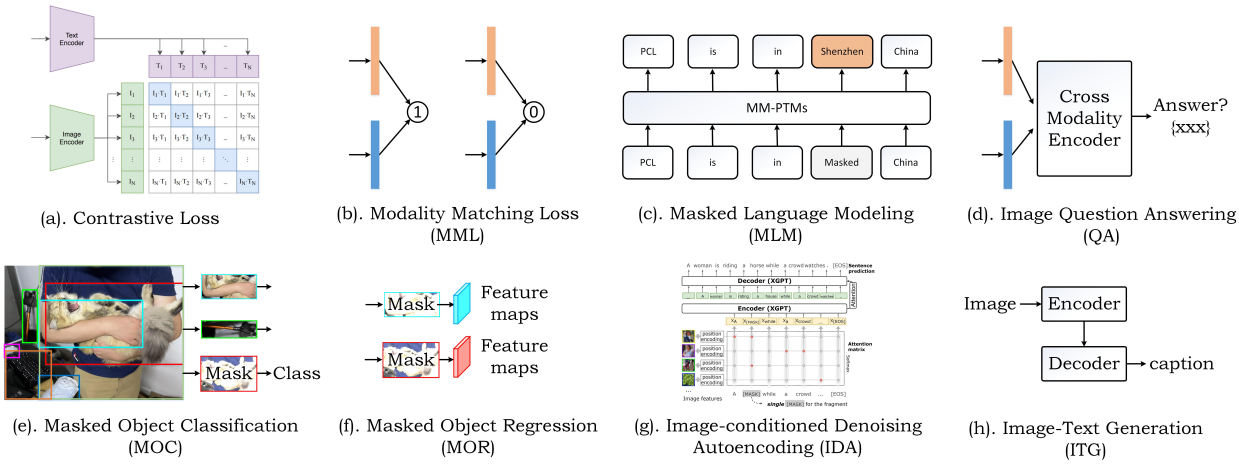
The contrastive losses in ALIGN are defined as follows:

$$\begin{aligned} \mathcal{L}_{i2t} &= -\frac{1}{N}\sum_{i}^{N} log\frac{exp(x_i^T y_i/\sigma)}{\sum_{j=1}^{N} exp(x_i^T y_j/\sigma)} \\ \mathcal{L}_{t2i} &= -\frac{1}{N}\sum_{i}^{N} log\frac{exp(y_i^T x_i/\sigma)}{\sum_{j=1}^{N} exp(y_i^T x_j/\sigma)} \\ \mathcal{L}_{CL} &= \mathcal{L}_{i2t} + \mathcal{L}_{t2i} \end{aligned} \tag{1}$$

where $\mathcal{L}_{i2t}, \mathcal{L}_{t2i}, \mathcal{L}_{CL}$ are an image-to-text classification loss function, a text-to-image classification loss function and the total contrastive loss respectively. The x_i is used to denote the normalized image embedding in the i-th pair, while the y_j denote the normalized embedding of text in the j-th pair. The N and σ are batch size and temperature parameter.

• **Modality Matching loss (MML)** is widely used in multi-modal pre-training big models due to the explicit or implicit alignment relationships between various modalities. For instance, Unicoder-VL [114] utilizes the Visual-linguistic Matching (VLM) for vision-language pre-training. They extract the positive and negative image-sentence pairs and train their model to predict whether the given sample pairs are aligned or not (in other words, to predict the matching scores). Different from regular negative image-text samples, the authors of InterBERT [115] design the image-text matching with hard negatives (i.e., ITM-hn) by selecting the highest TF-IDF similarities.

• **Masked Language Modeling (MLM)** is another widely pre-training objective, usually, the researchers usually mask and fill the input words randomly using special tokens. The surrounding words and corresponding image regions can be used as a reference for the masked word prediction. Wang et al. train SIMVLM [116] using the Prefix Language Modeling (PrefixLM), which executes the bi-directional attention on the prefix sequence and auto-regressive factorization on the rest tokens, respectively. The words are denoted as $w = \{x_1, \cdots, x_K\}$, and the image regions as $v = \{v_1, \cdots, v_T\}$. For MLM, the input words is masked as x_m by the mask indices m by generated randomly with a probability of $p\%$ The optimizing goal is to predict the masked words based on all image regions v and remaining words $x_{\neg m}$, by



(a). Contrastive Loss (b). Modality Matching Loss (MML) (c). Masked Language Modeling (MLM) (d). Image Question Answering (QA)

(e). Masked Object Classification (MOC) (f). Masked Object Regression (MOR) (g). Image-conditioned Denoising Autoencoding (IDA) (h). Image-Text Generation (ITG)

Fig. 5 Representative pre-training objectives used in MM-PTMs.

minimizing the negative log-likelihood:

$$\mathcal{L}_{MLM}(\theta) = -\mathbb{E}_{(x,v)} \log P_\theta(x_m | x_{\neg m}, v), \quad (2)$$

where θ is the trainable parameters. Beside MLM, PrefixLM in SIMVLM can also be adopted to pretrain vision-language representation:

$$\mathcal{L}_{PrefixLM}(\theta) = -\mathbb{E}_{\boldsymbol{x}\sim D} \log P_\theta(\boldsymbol{x}_{\geq T_p} | \boldsymbol{x}_{<T_p}), \quad (3)$$

where $\boldsymbol{x}$ is the given text sequence, D is the pre-training data and T_p is the length of a prefix sequence of tokens.

• **Masked Segment Modeling (MSM)** masks a continuous segment of given text using the special token, meanwhile, the MLM masks random words.

• **Image Question Answering (QA)** is used in LXMERT [117] to further expand the pre-training data, as many image-sentence pairs are image and question. The authors train their model to predict the answers as one of their pre-training objectives.

• **Masked Object Classification (MOC)** mainly focuses on masking the visual images using zero values. Then, people often take the predicted labels by object detector as the ground truth labels. This pre-training objective is widely used, such as Unicoder-VL [114]. Similar to MLM, the image regions can be masked by masking their viusal feature with a prabability of $p\%$. The goal is predict the object category of the masked image regions v_m^i. The encoder output of the masked image regions v_m^i is feed into an FC layer to predict the scores of T object classes, which further goes through a softmax function to be be transformed into a normalized distribution $g_\theta(v_m^i)$. The final objective is:

$$\mathcal{L}_{MOC}(\theta) = -\mathbb{E}_{(w,v)} \sum_{i=1}^{M} CE(c(v_m^i), g_\theta(v_m^i)), \quad (4)$$

where $c(v_m^i)$ is the ground-truth label.

• **Masked Object Regression (MOR)** is implemented to regress the masked feature or image regions. For example, the LXMERT [117] considers both MOC and MOR for their pre-training.

• **Image-Text Matching (ITM)** aims to align the image-text data. Negative training data is generated by randomly sampling, including negative sentences for each image, and negative images for each sentence. y is denoted by the gourd truth label for each image-text pair (v, t). A binary classification loss function is used for optimization:

$$\begin{aligned}\mathcal{L}_{ITM}(\theta) = -\mathbb{E}_{(v,t)}[&y \log s_\theta(v,t) \\ &+ (1-y)\log(1-s_\theta(v,t))],\end{aligned} \quad (5)$$

where s_θ is the image-text similarity score.

• **Unidirectional LM (UiDT)** Single direction history information is used for masked token prediction only, such as *left-to-right* and *right-to-left* language model objectives. Successful stories includes the ELMo [118], UNILM [119].

• **Bidirectional LM (BiDT)** Different from Unidirectional LM which predicts the masked token from a single direction only, the Bidirectional LM considers contextual information from both directions. Therefore, the contextual representations of text can be encoded more accurately. BERT [10], UNIML [119] and VLP [24] all adopt BiDT as one of their pre-training objective.

• **Sequence-to-Sequence LM (Seq2seq)** is a pre-training objective used in VLP [24], etc. It treats the inputs as different parts, each part can attend to different contexts.

• **Word-Region Alignment (WRA)** is used in UNITER [18] which target at explicitly achieves the fine-grained alignment between the multimodal inputs via Optimal Transport (OT) [120]. Specifically, the authors learn a transport plan which is a 2D matrix to optimize the alignment and resort to the IPOT algorithm [121] for approximate OT distance estimation. Then, the authors take this distance as the WRA loss to optimize their networks.

• **Action Prediction (AP)** target at evaluating whether the agent developed for vision-language navigation (VLN) can select the right actions based on the current image and instruction [122].

• **Image-conditioned Denoising Autoencoding (IDA)** is adopted in XGPT [11] to align the underlying image-text using an attention matrix. Even without the prior length of the masked fragment, the IDA could still reconstruct the whole sentence successfully.

• **Attribute Prediction (AttP)** is used to recover the masked tokens of attribute pairs, as indicated in ERNIE-ViL [123].

• **Relation Prediction (RelP)** is used in ERNIE-ViL [123] to predict the probability for each masked relation tokens to recover the masked relationship tokens.

• **Aligned Kaleido Patch Modeling (AKPM)** is proposed for the pre-training of Kaleido-BERT [124], which contains five kaleido sub-tasks, i.e., Rotation Recognition (RR), Jigsaw Puzzle Solving (JPS), Camouflage Prediction (CP), Grey-to-Color Modeling (G2CM), and Blank-to-Color Modeling (B2CM):

$$\begin{aligned}
\mathcal{L}_{RR} &= CE(y_r, \mathcal{F}(T,K,\theta)_{K_1_hidden}) \\
\mathcal{L}_{JPS} &= CE(y_j, \mathcal{F}(T,K,\theta)_{K_2_hidden}) \\
\mathcal{L}_{CP} &= CE(y_c, \mathcal{F}(T,K,\theta)_{K_3_hidden}) \\
\mathcal{L}_{G2CM} &= \sum KLD(k_{4i}, \mathcal{F}(T,K,\theta)_{K_4_hidden}) \\
\mathcal{L}_{B2CM} &= \sum KLD(k_{5i}, \mathcal{F}(T,K,\theta)_{K_5_hidden})
\end{aligned} \tag{6}$$

where CE represents the cross-entropy loss function, y_r denotes the rotation angle, K_p is the hidden output patch of size $p \times p$, KLD denotes the KL-divergence, and K_p are kaleido patches, among which k_{pi} is the masked out ones.

• **OBject Detection (OBD)** is introduced in the [125] as a direct set prediction to enhance the pre-training. Also, the authors consider object attribute prediction to learn the fine-grained semantic information. A negative log-likelihood loss is defined for OBD as follows:

$$\begin{aligned}
\hat{\sigma} &= \arg\min_{\sigma \in \phi_N} \sum_{i}^{N} \mathcal{L}_{match}(y_i, \hat{y}_{\sigma(i)}) \\
\mathcal{L}_{OBD}(y, \hat{y}) &= \sum_{i=1}^{N} [-log\hat{p}_{\hat{\sigma}(i)}(a_i) - log\hat{p}_{\hat{\sigma}(i)}(c_i) \\
&\quad + \mathcal{L}_{box}(b_i, \hat{b}_{\hat{\sigma}(i)}(i))]
\end{aligned} \tag{7}$$

where y denotes the ground truth set of objects and $\hat{y} = \{\hat{y}_i\}_{i=1}^{N}$, the number of elements is N, σ is the cost of a permutation of N elements, $\mathcal{L}_{match}(y_i, \hat{y}_{\sigma(i)})$ denotes the pair-wise matching loss between a prediction with index $\sigma(i)$ and ground truth y_i, $\hat{p}_{\hat{\sigma}(i)}(a_i), \hat{p}_{\hat{\sigma}(i)}(c_i)$ denotes the attribute and class probability, $\mathcal{L}_{box}(b_i, \hat{b}_{\hat{\sigma}(i)}(i))$ is a normalized loss of bounding box regression.

• **Image-Text Generation (ITG)** also plays an important role in the vision-language related pre-training tasks. The aligned image and text are capable of training a model for text generation based on a given image, for example, Xu et al. train the E2E-VLP [125] with ITG objective:

$$\mathcal{L}_{ITG} = -\sum_{(x,y)\in(\mathcal{X},\mathcal{Y})} log \prod_{t=1}^{n} P(y_t | y_{<t}, x) \tag{8}$$

where $\mathcal{X}$ represents the visual sequence with context, $\mathcal{Y}$ denotes the generated set of text, and the length of tokens in text y is n.

• **Video-Subtitle Matching (VSM)** considers two targets for the video-text pre-training task, i.e., (i) local alignment, (ii) global alignment, as used in HERO [126]. The score functions and the corresponding loss functions are defined as follows:

$$\begin{aligned}
&S_{local}(s_q, \boldsymbol{v}) = \boldsymbol{V}^{temp}\boldsymbol{q} \in \mathbb{R}^{N_v} \\
&S_{global}(s_q, \boldsymbol{v}) = max(\frac{\boldsymbol{V}^{temp}}{\|\boldsymbol{V}^{temp}\|}\frac{\boldsymbol{q}}{\|\boldsymbol{q}\|}) \\
&\mathcal{L}_h(S_{pos}, S_{neg}) = max(0, \delta + S_{pos} - S_{neg}) \\
&\mathcal{L}_{local} = -\mathbb{E}_D log(\boldsymbol{p}_{st}[y_{st}] + log(\boldsymbol{p}_{ed}[y_{ed}]) \\
&\mathcal{L}_{global} = -\mathbb{E}_D[\mathcal{L}_h(S_{global}(s_q, \boldsymbol{v}), S_{global}(\hat{s}_q, \boldsymbol{v})) \\
&\qquad + \mathcal{L}_h(S_{global}(s_q, \boldsymbol{v}), S_{global}(s_q, \hat{\boldsymbol{v}}))] \\
&\mathcal{L}_{VSM} = \lambda_1\mathcal{L}_{local} + \lambda_2\mathcal{L}_{global}
\end{aligned} \tag{9}$$

where s_q denotes the sampled query from all subtitle sentences, $\boldsymbol{v}$ is the whole video clip, $\boldsymbol{V}^{temp} \in \mathbb{R}^{N_v \times d}$ is the final visual frame representation generated by temporal transformer, $\boldsymbol{q} \in \mathbb{R}^d$ is the final query vector, $y_{st}, y_{ed} \in \{1, ..., N_v\}$ are the start and end index respectively, $\boldsymbol{p}_{st}, \boldsymbol{p}_{ed} \in \mathbb{R}^{N_v}$ represent probability vectors generated from the scores, $\boldsymbol{p}[y]$ indexes the y-th element of the vector $\boldsymbol{p}$, $\mathcal{L}_h$ denotes the combined hinge loss over positive and negative query-video pairs, $(s_q, \boldsymbol{v})$ is a positive pair while $(s_q, \hat{\boldsymbol{v}}), (\hat{s_q}, \boldsymbol{v})$ are negative ones replaced with one other sample in $\boldsymbol{v}$ and s_q respectively, δ is the margin hyper-parameter and λ_1, λ_2 are balancing factors.

• **Frame Order Modeling (FOM)** is treated as a classification problem in HERO [126], which targets reconstructing the timestamps of selected video frames. The objective of FOM is defined as follows:

$$\mathcal{L}_{FOM} = -\mathbb{E}_D \sum_{i=1}^{R} log\boldsymbol{P}[r_i, t_i] \tag{10}$$

where the number of reordered frames is R, $i \in [1, R], t_i \in \{1, ..., N_v\}$, r_i is the reorder index, $\boldsymbol{P} \in \mathbb{R}^{N_v \times N_v}$ is the probability matrix.

• **Textual Aspect-Opinion Extraction (AOE)** aims to extract aspect and opinion terms from the text, as noted in [127]. To handle the lack of label information required for supervised learning, the authors resort to other models for aspect extraction and opinion extraction. The obtained aspect and opinion terms are treated as labels for the AOE task.

• **Visual Aspect-Opinion Generation (AOG)** targets at generating the aspect-opinion pair detected from the input image [127].

• **Multimodal Sentiment Prediction (MSP)** enhance the pre-trained models by capturing the subjective information from vision-language inputs [127].

• **Modality-Level Masking (MoLM)** is used in [22] to learn the alignment among the text, vision, and audio. The authors mask out each modality independently with a certain probability.

• **Structural Knowledge Masking (SKM)** is proposed in [128] which attempts to mask the tokens selectively based on the cue provided by the knowledge entry. The masking probabilities is calculated to obtain mask indices M_w and M_r for each knowledge entry, the two items denote the words of sentences and visual regions of images need to be masked, respectively. The loss function of Structural Knowledge Masking Language Model can be formulated as:

$$\mathcal{L}_{SKMLM}(\theta) = -\mathbb{E}_{(W,R)\sim D} log P_\theta(\mathcal{W}_{M_w} | \mathcal{W}_{\setminus M_w}, \mathcal{R}_{\setminus M_r}) \tag{11}$$

where θ is the parameters. $\mathcal{W}_{\setminus M_w}$ and $\mathcal{R}_{\setminus M_r}$ represent the non-masked words of sequences and the remaining regions of images, respectively.

3.5 Pre-training Network Architecture

3.5.1 Self-attention and Transformer

In the large-scale pre-training era, most of current pre-trained models are inspired by the Transformer (which is mainly consisted of self-attention layers). It is originally developed for natural language processing tasks in 2017 [9] which sets new SOTA performance on many downstream tasks by a large margin. Such framework is also introduced into the computer vision community, therefore, the design of unified network architectures for various tasks and inputs is the current research hotspot.

Given the input x, an attention module A(x) is used to generate attention weights, then, some

procedures are conducted based on input x and A(x) to get the attended input x' = f(A(x), x). Many attention models are designed based on this idea, such as the channel attention, spatial attention, temporal attention, branch attention [129]. The self-attention scheme is a special case of attention mechanism, as shown in Fig. 6. More in detail,

$$Q, K, V = Linear(x) \tag{12}$$
$$A(x) = Softmax(QK) \tag{13}$$
$$f(A(x), x) = A(x)V \tag{14}$$

where the Linear denotes fully connected layers. On the basis of self-attention, the work mechanism of multi-head attention is the aggregation of parallel attention layers. Mathematically speaking,

$$MultiHead(Q, K, V) = [head_1, ..., head_h]W^O \tag{15}$$
$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V). \tag{16}$$

where [,] denotes the concatenate operation, W_i^Q, W_i^K, W_i^V and W^O are parameter matrices.

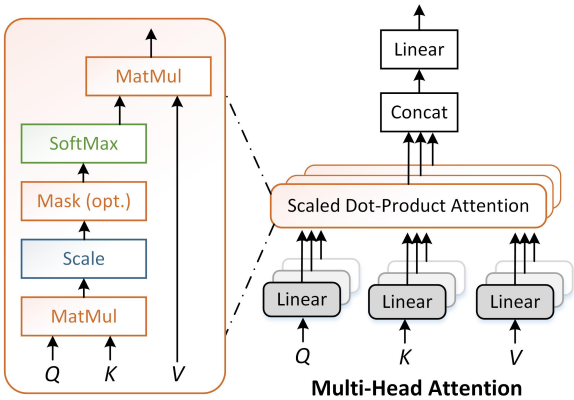


Fig. 6 An illustration of multi-head self-attention (MHSA) [9].

3.5.2 Single- and Multi-stream

The multi-layer transformer is widely used in many current MM-PTMs. The input of each modality is first extracted as feature embeddings by the independent encoder and then interacted with other modalities. According to the manner of multi-modal information fusion, two categories of MM-PTMs can be concluded, i.e., single- and cross-stream. In this subsection, we will present these two architectures separately.

• **Single-stream** Multi-modal inputs such as images and text are treated equally and fused in a unified model. The uni-modal features extracted from each modality are tokenized and concatenated by the separators as the input of the multi-modal transformer for multi-modal fusion, as shown in Fig. 8(a). In the transformer, the MHSA (multi-head self-attention) mechanism is usually adopted to interactively fuse the uni-modal features, then, the multi-modal fusion features are output from the class token of the transformer. Large-scale MM-PTMs based on single-stream structure includes VL PTMs (e.g., Oscar [17] and ALBEF [130]) and vision-language-audio pre-training model OPT [22]. Single-stream pre-training models perform token-level matching based on strong semantic correlation, e.g. object features of the image are matched with semantic features of object tags. It provides realistic interaction between uni-modal features, and multi-modal fusion features contain information from different modalities with better characterization capability.

• **Cross-stream** Features of different modalities are extracted in parallel by independent models and then are aligned by self-supervised contrastive learning in cross-stream architecture. The pre-training models obtain aligned uni-modal features rather than fused multi-modal features. As shown in Fig. 8(b), multi-modal fusion features are obtained by concatenating uni-modal features and fed into a MLP (Multi-Layer Perceptron) for pre-training objective learning. Representative large-scale MM-PTMs based on cross-stream structure include BriVL [106] and CLIP [77], etc. Compared with pre-training models based on single-stream, cross-stream models align different modality features into a consistent high-dimensional feature space, such as text semantics and visual image representation. Cross-stream pre-training models generally contain the CS pre-training objective and achieve embedding-level matching based on "weak semantic correlation" [106]. The structure of cross-stream models is more flexible, and modifying the branching structure of one modality of the model does not affect other modalities, making it easy to deploy in real scenarios. However, cross-stream models extract the aligned multi-modal common features, and how to effectively

exploit the information differences and complementarity between multi-modal data is an issue to be studied.

In addition, depending on the needs of the pre-training objectives, the structure of pre-training models can be divided into with and without a decoder. If pre-training objectives contain generative tasks, such as masked image reconstruction, generating matching images based on the text description, etc., the pre-training model adds a decoder after the encoder for converting multi-modal fusion features into the corresponding output.

3.5.3 Modality Interactive Learning

Most of current large-scale pre-trained multi-modal models adopt concatenate, add, Merge-attention, Co-attention, and Cross-attention [132] to achieve interactive learning between modalities. An introduction to these modules are given in the following paragraphs.

• **Merge-attention:** As shown in Fig. 7 (a), a unified feature representation is obtained by concatenating the input modalities. Then, this feature is fed into the fusion network. For example, the i-Code [131] flatten the visual inputs along the temporal and spatial dimensions. Note that the parameters of this attention model is shared by these input modalities.

• **Co-attention:** For the co-attention module, as shown in Fig. 7, each input modality has its own self-attention layers for modality-specific feature embedding. Then, the multiple embeddings are fused using a cross-attention layer.

• **Cross-attention:** For the multi-modal task, the key step is how to design a fusion module to connect the multi-modality inputs effectively. For instance, the cross-attention layer is proposed by Suo et al. [132], which integrate the image and language subtly for visual question answering. Specifically, they mutually input one modality into the Q-branch of another self-attention network. Then, the output of two modalities are concatenated as one unified representation for final prediction.

• **Tangled-transformer:** The TaNgled Transformer (TNT) [133] is proposed to handle the action-, regional object-, and linguistic-features, simultaneously, using three Transformer modules. As shown in Fig. 7 (d), the authors inject one modality to the Transformer network designed for other modality to enhance the interactions.

• **Inter-Modality Contrastive Learning:** The contrastive learning is widely used for inter-modality relation modelling, such as the CLIP [77] and its following-up works [19, 104, 134–138]. The representative work SCALE [104] is trained with Self-harmonized Inter-Modality Contrastive Learning (SIMCL), which can be written as:

$$\mathcal{L}_{CL}(d_i^{(0)}, d_i^{(1)}) = -log \frac{exp(Sim(f_i^{(0)}, f_i^{(1)})/\tau)}{\sum_{m=0}^{1}\sum_{k=1}^{N} \mathbf{1}_{[k\neq i]} exp(Sim(f_i^{(m)}, f_k^{(1-m)})/\tau))}, \tag{17}$$

where $(d_i^{(0)}, d_i^{(1)})$ is a positive pair, and the pairing of $d_i^{(0)}$ and other samples will bring us negative training data. $f_i^{(0)}, f_i^{(1)}$ are feature embedding of $(d_i^{(0)}, d_i^{(1)})$ respectively. The Sim denotes the cosine similarity, $\mathbf{1}_{[k\neq i]}$ is the binary indicator function, τ is a temperature parameter.

3.6 Pre-training using Knowledge

Conventional pre-trained models suffer from poor logical reasoning and lack of interpretability. To alleviate those problems, it is straightforward to involve knowledge, deep understanding of data, in pre-training models, i.e., pre-training using knowledge also known as Knowledge Enhanced Pre-Trained Models (KEPTMs) shown in Fig. 9.

Knowledge Representation Learning By learning to represent symbolic knowledge, usually in the form of entities and relations, knowledge representation learning enables neural network based models to fuse knowledge and improve their reasoning capabilities. Similarity-based models and graph neural network (GNN) models are two major methods of knowledge representation learning.

• **Similarity-based Models** Given similarity-based scoring functions, similarity-based models measure the similarity of latent semantics between two entities. Translation-based models are representatives of similarity-based models, as the distance in the vector space is often used to describe the similarity. TransE firstly models relations by translations, which operates on entity embeddings at low-dimension [197]. To deal with mapping properties of relations efficiently in complex models, such as reflexive, one-to-many, many-to-one and many-to-many,

Table 3 The summary of mainstream multi-modal pre-trained big models (Part-I).

No.	Model	Pub.	Modality	Architecture	Objective	Highlights	Parameters	Code
01	VisualBERT [139]	arXiv-2019	image-text	Trans, BERT	GR, MML	A simple and strong baseline for VLP	170M	URL
02	ViLBERT [140]	NeurIPS-2019	image-text	Trans	CS, GR	First adopt co-attention for MM pre-training	274M	URL
03	LXMERT [117]	EMNLP-2019	image-text	Trans	QA, MOR, MOC, MML, MLM	Propose a cross-modality encoder for vision-language pre-training	183M	URL
04	B2T2 [141]	EMNLP-2019	image-text	ResNet, BERT	MML, GR	Embed bounding box into text transformer in a early fusion manner	-	URL
05	Unicoder-VL [114]	AAAI-2020	image-text	Trans	GR, MML, MOC	Single transformer encoder for VLP	170M	URL
06	VL-BERT [142]	ICLR-2019	image-text	BERT	GR, MOC	MM PTMs and faster rcnn are jointly trained	-	URL
07	VLP [143]	AAAI-2020	image-text	Trans	BiDT, Seq2seq	Unified encoder-decoder network architecture	-	URL
08	UNITER [18]	ECCV-2020	image-text	Trans	MRA, MML	Propose an OT-based Word-Region Alignment objective	110M	URL
09	12-IN-1 [144]	CVPR-2020	image-text	Trans	CS, GR	Training jointly on 12 different datasets in a multi-task learning manner	270M	URL
10	VisDial-BERT [145]	ECCV-2020	image-text	Trans	MLM, NSP, MIR	Pre-training on image-text corpus and finetuning on visual dialog	-	URL
11	ImageBERT [87]	arXiv-2020	image-text	Trans	MOC, MLM, MML, MOR	Indicating that multi-stage pre-training works better	170M	-
12	PREVALENT [122]	CVPR-2020	image-text	Trans	MLM, AP	Pre-training for vision and language navigation	-	URL
13	XGPT [11]	NLPCC-2021	image-text	Trans	IC, MLM, IDA, MOR	Novel IDA pre-training; Share parameters between encoder and decoder	-	-
14	InterBERT [115]	arXiv-2020	image-text	Trans	MSM, MOC, ITM-hn	Finding that all-attention works better than co-attention for modal interaction	173M	URL
15	PixelBERT [20]	arXiv-2020	image-text	CNN, Trans	MLM, MML	First to align vision and language in pixel and text-level	142M	-
16	OSCAR [17]	ECCV-2020	image-text	Trans	CS, MLM	Align the visual patches with word embeddings by using object tags as anchor points	155M	URL
17	pyramidCLIP [146]	arXiv-2022	image-text	CNN+Trans	CS	Hierarchical image-text contrastive learning	-	-
18	FashionBERT [147]	RDIR-2020	image-text	BERT	MLM, MOR, MML	Use image patches for fashion domain instead of RoIs	-	URL
19	VILLA [148]	NeurIPS-2020	image-text	Trans	MLM, MOR, MML	Pre-training with adversarial learning	-	URL
20	ERNIE-ViL [123]	AAAI-2021	image-text	Trans	MOC, AttP, RelP, MLM, MOR, MML	Use the knowledge obtained from scene graph	-	URL
21	KVL-BERT [149]	KBS-2021	image-text	BERT	MOC, MLM	Integrate commonsense knowledge for visual commonsense reasoning	-	-
22	VinVL [113]	CVPR-2021	image-text	Trans	MTL, 3-way CS	Verifying that visual feature matters in VLP, i.e., strong object detector brings better results	157M	URL
23	VL-T5 [150]	ICML-2021	image-text	Trans	MLM, VQA, MML, VG, GC	Unified framework for VL via generating texts	400M	URL
24	ViLT [151]	ICML-2021	image-text	Trans	MLM, MML	Use linear embedding only for Fast VL transformer	87M	URL
25	ALIGN [21]	ICML-2021	image-text	EfficientNet, BERT	CS	Milestone for image-text pre-training using noisy data	300M	-
26	Kaleido-BERT [124]	CVPR-2021	image-text	Trans	MLM, MML, AKPM	Use saliency detector to generate multi-grained patches	-	URL
27	MDETR [152]	ICCV-2021	image-text	CNN+Trans	STP, MML	A text-modulated detection system which can be trained in an end to end way	-	URL
28	SOHO [153]	CVPR-2021	image-text	CNN+Trans	MLM, MOR, MML	Use a dynamic-updated visual dictionary for vision-language alignment	-	URL
29	E2E-VLP [125]	ACL-2021	image-text	Trans	OBD, ITG	The first PTM for vision-language understanding and generation	94M	-
30	PIM [154]	NeurIPS-2021	image-text	Trans	MLM, MML, MOR	Measure and reveal the V+L fusion using the proposed inter-modality flow metric	48M	-
31	CLIP-ViL$_p$ [137]	arXiv-2021	image-text	Trans	MLM, VQA, MML	Take the CLIP visual encoder as its visual backbone	-	URL
32	ALBEF [130]	NeurIPS-2021	image-text	Trans	CS, GR	Design a momentum model to address noisy data	210M	URL
33	SimVLM [116]	arXiv-2021	image-text	Trans	PrefixLM	Simple VL model using single PrefixLM pre-training objective only	-	-
34	MURAL [155]	arXiv-2021	image-text	Trans	CS	Adopt multi-task contrastive learning objective (image-text, text-text)	430M	-
35	VLMo [156]	arXiv-2021	image-text	Trans	MLM, MML, CS	Jointly learns visual-, text-encoder and a fusion encoder	-	URL

Table 4 The summary of mainstream multi-modal pre-trained big models (Part-II).

No.	Model	Pub.	Modality	Architecture	Objective	Highlights	Params	Code
36	METER [157]	CVPR-2022	image-text	Trans	MLM, MOR, MOC, MML	An empirical study on VLP	-	URL
37	VideoBERT [158]	ICCV-2019	video-text	BERT	MLM	A simple model for video-text feature learning	-	URL
38	CBT [159]	arXiv-2019	video-text	Trans	NCE	Self-supervised contrastive bidirectional Transformer	15M	-
39	UniVL [160]	arXiv-2020	video-text	Trans	MLM, MFM, MML, ITG	A unified model for multimodal understanding and generation	-	URL
40	HERO [126]	EMNLP-2020	video-text	Trans	MLM, MFM, VSM, FOM	Hierarchical Transformer-based model trained with newly proposed VSM and FOM	-	URL
41	MMFT-BERT [161]	EMNLP-2020	image-text	BERT	Classification	Adopt multiModal fusion Transformer for modality fusion	-	URL
42	ActBERT [133]	CVPR-2020	image-text	Trans	CS, GR	Extract actions explicitly as one of the inputs	-	-
43	CLIP [77]	ICML-2021	image-text	Resnet, Trans	CS	Milestone for image-text pre-training using noisy data	88.6M	URL
44	Frozen [92]	ICCV-2021	video/image-text	Trans	MML	Jointly optimize the model on both images and videos	180.4M	URL
45	RegionLearner [162]	arXiv-2021	video-text	Trans	MML	Implicitly learning object region without position supervision	-	URL
46	UNIMO [163]	arXiv-2020	image-text	Trans	CS	Adapt to single-, multi-modal understanding and generation tasks effectively	-	URL
47	DALL-E [164]	ICML-2021	image-text	Trans	ELB	Achieve high quality image generation without using any of the training labels	12B	URL
48	BriVL [106]	arXiv-2021	image-text	Trans	InfoNCE	The first Chinese large-scale MM-PTMs	10B	URL
49	VLC [165]	arXiv-2022	image-text	ViT	MIM, MLM ITM	Built on top of MAE that does not require trained on ImageNet	87M	URL
50	M6 [103]	arXiv-2021	image-text	Trans	LM	The largest pretrained model in Chinese	100B	-
51	CogView [166]	NeurIPS-2021	image-text	Trans	NLL	The first open-source large text-to-image transformer	4B	URL
52	VATT [167]	NeurIPS-2021	Video, Audio, Text	Trans	NCE, MIL-NCE	Modality-specific or Modality-agnostic triplet modality pre-trained model	306.1M	URL
53	OPT [22]	arXiv-2021	image, Audio, Text	Trans	MLM, MVM, MoLM MAM, DTR, DIR	The first model pre-trained using triplet modalities	-	-
54	Florence [168]	arXiv-2021	image-text	CoSwin	UniCL	Multi-dimensional expansion of representations	893M	-
55	ROSITA [128]	MM-2021	image-text	Trans	SKM, MLM, MRM	Fuse the intra-, cross-modality knowledge, and SKM	-	-
56	VLCDoC [169]	arXiv-2022	image-text	Trans	CS	Contrastive Pre-Training for document classification	-	-
57	MVP [170]	arXiv-2022	image-text	ViT	MIM	Multimodality-guided visual pre-training leads to impressive gains	-	-
58	GilBERT [171]	IR-2021	image-text	BERT	MLM, MOR	Considers both realistic and synthetic data for VLP	-	-
59	COTS [172]	arXiv-2022	image-text	Trans	CS, KLD, MVLM	Token- and task-level interaction are proposed to enhance cross-modal interaction	-	-
60	U-VisualBERT [173]	NAACL-2021	image-text	Trans, BERT	GR, MML	*Unpaired* image-text data for pre-training	-	URL
61	Flamingo [174]	arXiv-2022	image-text	NFNet	CS	Pre-training on interleaved visual and text data as input	80B	URL
62	M3P [175]	CVPR-2021	image-text	BERT	xMLM, MC-MLM, MC-MRM	Multitask, Multilingual, Multimodal Pre-training	-	URL
63	BLIP [176]	arXiv-2022	image-text	BERT	CS, MML, MLM	Propose the multimodal mixture of encoder-decoder, and captioning-filtering scheme	224M	URL
64	NUWA [177]	arXiv-2021	image-text	Trans	T2I, T2V, V2V	A 3D transformer framework can handle image, text, and video, simultaneously	809M	URL
65	TCL [178]	CVPR-2022	image-text	BERT	CMA, IMC, LMI ITM, MLM	The first work considers local structure information for multi-modality representation learning	123.7M	URL
66	SCALE [179]	CVPR-2022	image, text, table video, audio	BERT	MRP, MLM, MEM MFP, MFP, MAM	A unified model to handle five modalities	-	URL
67	Clinical-BERT [180]	AAAI-2022	image-text	BERT	CD, MMM MLM, IMM	The first work to learn domain knowledge during pre-training for the medical domain	102M	-
68	RegionCLIP [181]	CVPR-2022	image-text	Trans	Distillation loss, CS	Learn region-level visual representations based on CLIP	-	URL
69	ProbES [182]	ACL-2022	image-text	LSTM, ViLBERT	Ranking loss	Prompt-based learning for VLN based on CLIP	-	URL
70	GLIP [183]	CVPR-2022	image-text	BERT	CS	Unifying the object detection and grounding into a unified framework	394M	URL

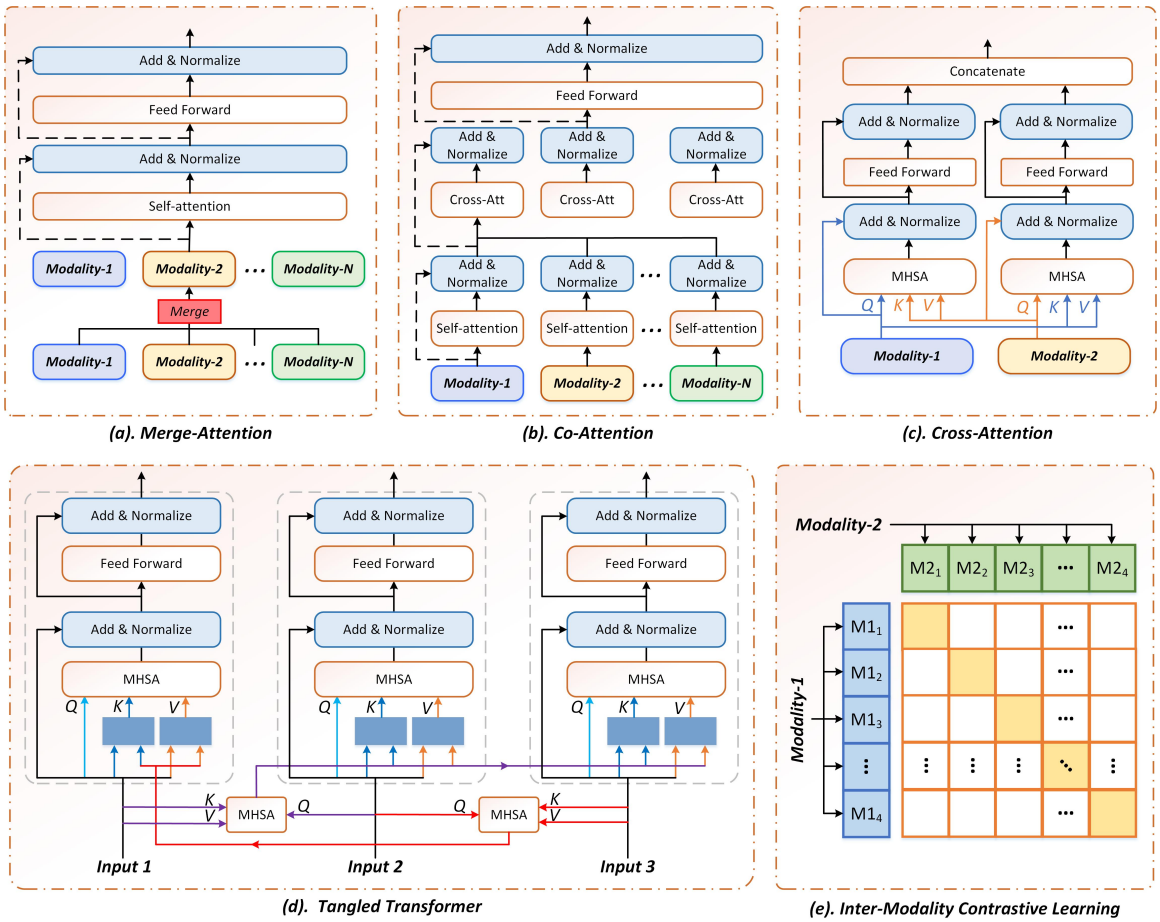


Fig. 7 The widely used modality interactive learning modules for MM-PTMs. (a) Merge-attention [131], (b) Co-attention [131], (c) Cross-attention [132], (d) Tangled-transformer [133], and (e) Contrastive learning [77].

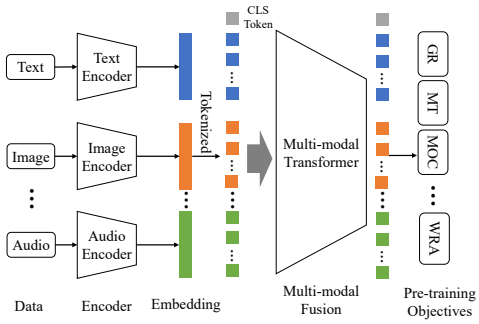
TransH is proposed to model a relation as a translation operation on a hyperplane [198]. TransR is proposed to embed entity and relation in a separated spaces to capture different aspects of entities over various relations [199]. Compared with TransR, not only the diversity of relations but also entities are considered in TransD [200]. To deal with heterogeneity and imbalance issues brought by knowledge graphs but ignored by aforementioned translation-based models, transfer matrices are replaced with adaptive sparse matrices in TranSparse, because the number of entities linked by relations determines sparse degrees [201]. Besides translation-based models, tensor or matrix factorization approaches have also been proposed for multi-relational data by introducing scoring or ranking functions to measure how likely the semantic matching is correct. With the latent components, RESCAL is capable of collective learning and can provide an efficient algorithm of the factorization of a three-way tensor [202]. NTN introduces an expressive neural tensor network for reasoning over relationships between two entities [203]. DistMult presents a general framework for multi-relational learning and shows the effectiveness of a simple bilinear formulation [204]. SME designs a new neural network architecture to encode multi-relational graphs or tensors into a flexible continuous vector space, so that multi-relational semantics can be learnt [205]. HolE is proposed to learn compositional vector space representations of entire knowledge graphs by employing holographic models of associative memory and circular correlation to create compositional representations [206].

• **Graph Neural Network Models** To further leverage the structure of the graph rather than collections of triplets, graph neural network

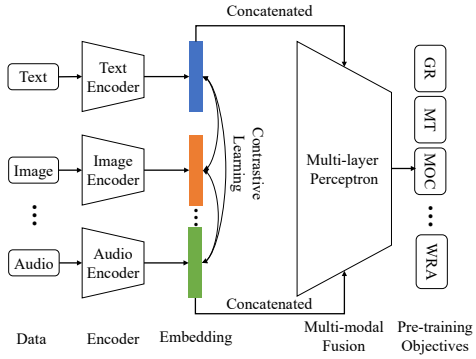
Table 5 The summary of mainstream multi-modal pre-trained big models (Part-III).

No.	Model	Pub.	Modality	Architecture	Objective	Highlights	Parameters	Code
71	VLP-MABSA [127]	ACL-2022	image-text	BERT	MLM, AOE, MRM AOG, MSP	Task-specific VL-PTMs for multimodal aspect-based sentiment analysis	-	URL
72	R2D2 [184]	arXiv-2022	image-text	ViT, BERT	GCPR, FGR, MLM	A two-way distillation strategy is proposed, i.e., target- and feature-guided distillation	-	-
73	DeCLIP [19]	ICLR-2022	image-text	ViT	InfoNCE, SS MVS, NNS	Learn generic visual features in a data efficient way	276M	URL
74	DeFILIP [136]	arXiv-2022	image-text	ViT, ResNet	CS	A benchmark for CLIP and its variants	-	URL
75	SLIP [185]	arXiv-2021	image-text	ViT	CS, InfoNCE	Combine the self-supervised learning and CLIP pre-training in a multi-task framework	38M	URL
76	FILIP [186]	arXiv-2021	image-text	ViT	CS	Cross-modal interactive learning for finer-level alignment	-	-
77	SemVLP [187]	arXiv-2021	image-text	Trans	MLM, MOP, ITM, QA	Fuse the single- and two-stream architectures	2.1B	-
78	CoCa [188]	arXiv-2022	image-text	Trans	CS, ITG	Jointly pre-train image text model with contrastive loss and captioning loss	-	-
79	HiVLP [189]	arXiv-2022	image-text	Trans	LRM, HRL, VLM	Accelerate image-text retrieval via hierarchical retrieval	-	-
80	CLIP-Event [135]	CVPR-2022	image-text	Trans	CS	Consider event structural knowledge and prompts in the pre-training phase.	-	URL
81	AudioCLIP [190]	ICASSP-2022	image-text-audio	Trans	CS	Build a triplet modality based PTMs like CLIP	30M	URL
82	VL-BEiT [191]	arXiv-2022	image-text	Trans	MLM, MIM, MVLM	Share the Transformer network on both monomodal- and multimodal-data	-	URL
83	MV-GPT [192]	arXiv-2022	image-text	BERT	MLM, LG	Pre-train both a multi-modal video encoder and a sentence decoder jointly.	117M	-
84	MMKD [193]	arXiv-2022	image-text	BERT	ITM	Iteratively execute knowledge discovery and model pre-training for continuous learning	-	-
85	GLIPv2 [194]	arXiv-2022	image-text	Swin, BERT	PGL, CS, MLM	Serves both the localization and understanding tasks.	-	URL
86	LIMoE [195]	arXiv-2022	image-text	Trans	CS	multi-modal pre-training with a sparse mixture of experts model	675M	-
87	VLMixer [196]	arXiv-2022	image-text	Trans	MLM, CMCL, MTM	Implicit cross-modal alignment learning in unpaired VLP.	-	URL
88	ProtoCLIP [138]	arXiv-2022	image-text	Trans	CS	Combine the CLIP loss and prototypical supervisions for VLP.	-	URL
89	i-Code [131]	arXiv-2022	image-text-audio	Trans	MLM, MVM MSM, CS	It can handle different combinations of modalities (such as single-, dual-, and triple-modality) into a single representation space.	906M	-

models are employed to embed entities and relations. As convolutional neural networks (CNNs) are extremely efficient architectures in recognition tasks over different domains, they are generalized to graphs based on hierarchical clustering of the domain and the spectrum of the graph Laplacian in [207]. Inspired by the pioneering work, further efforts have been done on graph convolutional networks (GCNs), such as semi-supervised classification [208], unsupervised learning based on the variational auto-encoder (VAE) [209], inductive representation learning to sample and aggregate features from a node's local neighborhood [210], and attention mechanism by leveraging masked self-attentional layers [211]. Beyond GCNs, R-GCNs is developed to deal with the highly multi-relation data characteristic of realistic knowledge bases [212]. A structure-aware convolutional network (SACN) takes the benefit of GCN and ConvE [213] together, where GCN as the encoder utilizes knowledge graph node structure and ConvE as the decoder enables the translational feature [214]. To further enhance Graph Attention Networks (GATs) and capture both entity and relation features within any entity's neighborhood, another model is proposed for attention-based feature embedding [215]. To leverage various composition operations for embedding entities



(a) Architecture of single-stream pre-training multi-modal model



(b) Architecture of Cross-stream pre-training multi-modal model

Fig. 8 Pre-training network architecture.

and relations in KGs and ever-increasing number of relations, a composition-based GCN named CompGCN is proposed to embed both nodes and relations jointly [216].

Knowledge Fusion Methods How to fuse knowledge into pre-trained models and improve their logical understanding of data after knowledge representation learning remains a challenge to researchers. According to the category of knowledge provided, KEPTMs roughly contain two categories: unstructured knowledge and structured knowledge enhanced pre-trained models.

- **Unstructured KEPTMs** Unstructured knowledge often refers to the knowledge without structures involved, which is in the form of plain text, like the words or phrases. Although some literatures introduce entities as supervised data and achieve promising performance, structural information is ignored while only entities are used to enable PTMs to learn semantics or attain extra key features from them. Word-aligned attention aligns the character-level attention to the word level to exploit explicit word information in Chinese [217]. SentiLARE also introduces part-of-speech tag and sentiment polarity to build word-level linguistic knowledge [218]. As unstructured text trained neural language models can store knowledge implicitly, PTMs can be further fine-tuned to explicitly retrieve knowledge without access to external knowledge or context [219].
- **Structured KEPTMs** Contrary to unstructured KEPTMs, structured KEPTMs take account of sorts of structural information, including syntax-tree, rules and knowledge graphs. Syntax-BERT incorporates syntax trees effectively and efficiently into pre-trained Transformers [220]. LIMIT-BERT learns language representations across multiple linguistics tasks including constituent and dependency syntactic parsing [221]. Syntax-GNN is proposed to learn syntax representations by using dependency trees and fusing the embeddings into transformers [220]. Knowledge graphs (KGs) provide structural knowledge in the form of entities and relations between them. An enhanced language representation model ERNIE is trained by utilizing both large-scale textual corpora and knowledge graphs, so that it can simultaneously leverage lexical, syntactic and knowledge [222]. Similar work named KnowBert is also proposed for large-scale models to embed multiple knowledge bases with entity linkers, which retrieves relevant entity embeddings and updates contextual word representations by the word-to-entity attention [223]. Moreover, the reasoning capability is also developed by finding supporting-facts, based on a large external knowledge base [224, 225]. Rules, in the form of constraints or even logical expressions, are preferred due to their interpretability and accountability. HEX graphs are proposed to enhance existing models by capturing semantic relations between labels applied to the same object [226].

Knowledge Evaluation Tasks Besides conventional performance metrics, more knowledge-oriented tasks are required to evaluate the capability of KEPTMs and inspect whether external knowledge really helps models understand data semantically. Knowledge evaluation tasks are severed as testbeds to ensure the effectiveness of knowledge fusion methods. Currently, knowledge

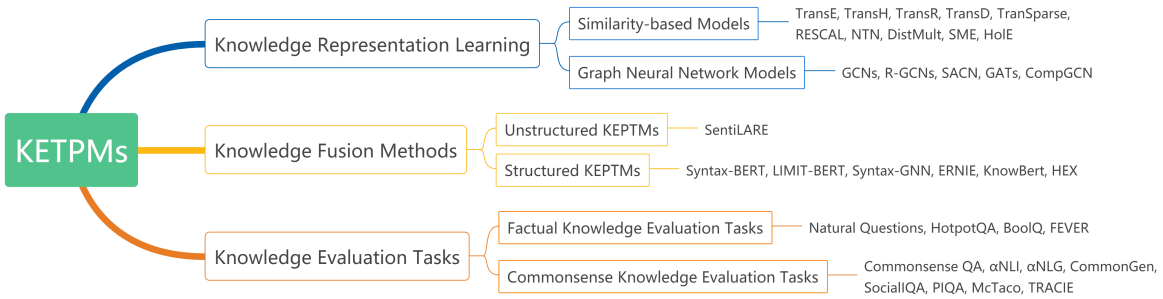


Fig. 9 The taxonomy of Knowledge Enhanced Pre-Trained Models (KEPTMs).

evaluation tasks mainly focus on NLP tasks and can be categorized into two groups based on the types of required knowledge: factual knowledge and commonsense knowledge evaluation tasks.

• **Factual Knowledge Evaluation Tasks** Factual knowledge is the knowledge of facts, including specific details and elements to describe the objective facts [28]. Factual knowledge evaluation tasks focus on testing models' reasoning ability on factual knowledge over various domains, like answering questions by giving a fact or judging the correctness of a given fact. Natural Questions is the first large publicly available dataset and robust metrics are also introduced to evaluate the performance of question answering (QA) systems [227]. HotpotQA, another QA dataset, provides supporting facts at sentence-level for reasoning and new factoid comparison questions [228]. Different from the above two open-domain QA tasks, BoolQ only involves yes/no naturally occurring questions, namely verifying facts generated in unprompted and unconstrained settings, but those queries involve with complicated and non-factoid information so that make it unexpectedly challenging [229]. Another fact extraction and verification task FEVER is proposed and a new type of claims NotEnoughInfo is introduced beside Supported and Refuted [230]. Entity linking, linking entities from a knowledge base to the corresponding textual mentions in a corpus, can also evaluate how well a model understands the factual knowledge [231].

• **Commonsense Knowledge Evaluation Tasks** Commonsense knowledge refers to the information generally accepted by the majority of people concerning everyday life, i.e. the practical knowledge about how the world works [29]. Like factual knowledge evaluation tasks, Commonsense QA also focuses on QA, but such QA requires prior knowledge outside the given document or context [232]. To extend the QA task Abductive Natural Language Inference (αNLI), Abductive Natural Language Generation (αNLG), a conditional generation task, is also proposed to explain given observations in natural language [233]. CommonGen further explicitly tests models for the ability of generative commonsense reasoning due to its rigorous requirements on both relation reasoning and compositional generalization [234]. Besides general commonsense evaluation tasks evaluating how well models understand daily scenarios, specific commonsense knowledge ones are further designed for different scenarios. SocialIQA, a large-scale benchmark for social commonsense reasoning, is challenging even for PTMs [235]. Beside human interactions, physical interactions are also important in commonsense knowledge, hence the task of PIQA is introduced for physical commonsense reasoning [236]. Temporal commonsense is crucial for understanding the timing of events, for example duration, frequency, and order, leading to correct reasoning. McTaco defines five classes of temporal commonsense [237], while TRACIE evaluates models' temporal understanding of implicit events [238].

3.7 Characteristics of Different Pre-trained Big Models

In the aforementioned paragraphs, we give a review to the main streams of multi-modal pre-trained models and highlight the features of each model in Table 3, Table 4, and Table 5. In this subsection, we compare and analyze the characteristics of these models. Specifically, the early

multi-modal pre-trained big models usually design an interactive learning module, for example, the ViLBERT [140], LXMERT [117]. They integrate the co-attention or cross-attention mechanism into their framework to boost the feature representation between multiple inputs. Actually, these models obey the idea of interactive fusion of traditional small models. This allows for seamless integration with numerous downstream tasks and providing a high degree of flexibility. In contrast, many current big models directly process the inputs using projection layers and feed them into a unified network like the Transformers, including Unicoder-VL [114], VideoBERT [158], UniVL [160]. More and more works demonstrate that the powerful Transformer network can achieve comparable or event better performance.

There are also some works make full use of existing big models and carry out secondary development to achieve a higher performance [181, 190]. To address the issues caused by shortage of paired multi-modal data, some researchers propose to training their model using unpaird data [173]. These models show the great potential of processing massive multi-modal data. Unlike general big models, some models are specifically designed for a specific task or domain, like the e-commerce, or Indoor navigation. This provides conditions and convenience for fully mining more detailed domain knowledge assist the pre-training process.

4 Downstream Tasks

After the pre-training phase, the researchers usually test their model on many downstream tasks to validate the powerful ability. Specifically, the generative tasks, classification tasks, regression tasks are adopted for the validation which will be discussed below. As a new learning paradigm, the prompt learning which target at modifying the downstream tasks to fit the pre-trained big model draws more and more attention. In this part, several representative prompt learning algorithms are also reviewed. An overview of these downstream tasks are visualized in Fig. 10.

4.1 Generative Tasks

Image/Video Captioning attempt to describe content of input image or video using a couple of sentences. Usually, a visual encoder is used to encode the input image/video, then, a language decoder is adopted for sentence prediction in a word by word manner. NoCaps [239] is proposed by Agrawal et al. in 2019. It is also an image captioning task but focus on developing generalized captioning models.

Visual Dialogue (VD) attempt to let the AI agent to talk with humans by holding a meaningful dialog about the visual content [240].

Multi-modal Machine Translation (MMT) is a task that targets translating the source sentence into a different language based on the paired image [241].

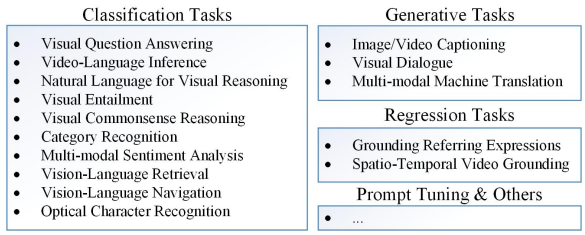


Fig. 10 An overview of downstream tasks reviewed in this paper.

4.2 Classification Tasks

Visual Question Answering (VQA) model is provided with an image and a question, and asked to produce an answer [242]. The relations between GQA [86] and VQA is similar to the NoCaps and the standard captioning task. It is introduced to address key drawbacks of previous VQA datasets, and generate novel and diverse questions from a robust question engine, which sufficiently considers the content and structure.

Video-Language Inference (VLI) is proposed by Liu et al. [243] in year 2020, which aims at understanding the video and text multimodal data.

Natural Language for Visual Reasoning (NLVR) can be seen as a binary classification problem. As noted in [244], the model needs to judge the authenticity of a statement for the image.

Visual Entailment (VE) [245] is a triplet-label classification problem derived from Text Entailment (TE) task [246]. The VE model needs to predict whether the given image semantically entails the text. The three labels are *entailment*, *neutral* or *contradiction*.

Visual Commonsense Reasoning (VCR) [247] is a variation of VQA, which require a machine to provide a rationale justification and answer correctly for the given challenging problem.

Category Recognition (CR) is a classification problem which attempt to predict the category of given image. Many computer vision tasks are belong to this downstream task, such as pedestrian attribute recognition [248], action recognition [134].

Multi-modal Sentiment Analysis (MSA) is a multi-modal fusion task proposed for sentiment analysis [249], which attempt to aggregate various homogeneous and/or heterogeneous modalities for more accurate reason. The modalities can be text, visual and acoustic, etc.

Vision-Language Retrieval (VLR) can be used in many applications, such as text-based person search [250], or general object retrieval based on language [251].

Vision-Language Navigation (VLN) [252, 253] is task that the agents learn to navigate in 3D indoor environments following the given natural language instruction. A benchmark for the popular VLN can be found at the following leaderboard.

Optical Character Recognition (OCR) target at convert the images of Diverse text information into machine-encoded text. Usually, the OCR system contains both text detection and text recognition modules.

4.3 Regression Tasks

Grounding Referring Expressions (GRE) takes the visual image and language description as input, and output the location of target object described by the language [254–256]. Similar tasks defined on videos are termed **Spatio-Temporal Video Grounding (STVG)** [257] or **Tracking by Natural Language** [258–260].

4.4 Prompt Learning

To make full use of pre-trained big models, the prompt learning (also called prompt tuning) is proposed to re-formulate the downstream tasks to fit the objectives of pre-trained models, including CPT [261], CPL [262]. Also, some prompt tuning schemes are designed to fix the parameters of the large model and adjust the parameters as little as possible to achieve good results, such as the VPT [263], CoOp [264], CoCoOp [265]. To be specific, the VPT [263] fixes the parameters of ViT models and integrates the prompt vectors as additional input. It achieves good performance even only tune the parameters of classification head and prompts. CoOp [264] achieves huge improvements by tuning the context words into a set of learnable prompt vectors. Conditional Context Optimization (CoCoOp) [265] is developed based on CoOp which learns an external network to generate input-conditional tokens for each image. It addresses the issue of class shift significantly using such dynamic prompts.

5 Experimental Analysis

Considering the complexity and numbers of MM-PTMs, it is almost impossible to reproduce pre-training tasks in a short amount of time. Therefore, the experiments and related analyses of the pre-training are ignored in this paper. However, we still want to summarize a more complete review paper for the readers, thus, we extract the experimental results of the corresponding downstream tasks from their paper and compare them to the shared benchmark datasets. More detailed results can be found in Table 3 and Table 4.

5.1 Model Parameters and Training Information

As shown in Fig. 11 (a), the large-scale MM-PTMs are emerging in the year 2019 and the number of papers shows an increasing trend year by year [5]. From the Fig. 11 (b), it is easy to find that current large-scale PTMs are optimized on servers with more than 8 GPUs. Also, many of them are trained using more than 100 GPUs, such as BriVL (128) [106], VLC (128) [165], M6 (128) [103], SimVLM (512) [116], MURAL (512) [155], CLIP (256) [19], VATT (256) [167], Florence (512) [168], FILIP (192) [186]. Some MM-PTMs are trained on TPUs with massive chips, for example, the largest model of Flamingo [174] is trained for 15 days on 1536 chips. From all these cases, we can see the

[5] Note that only half a year's results (the year 2022, from January to June) have been counted.

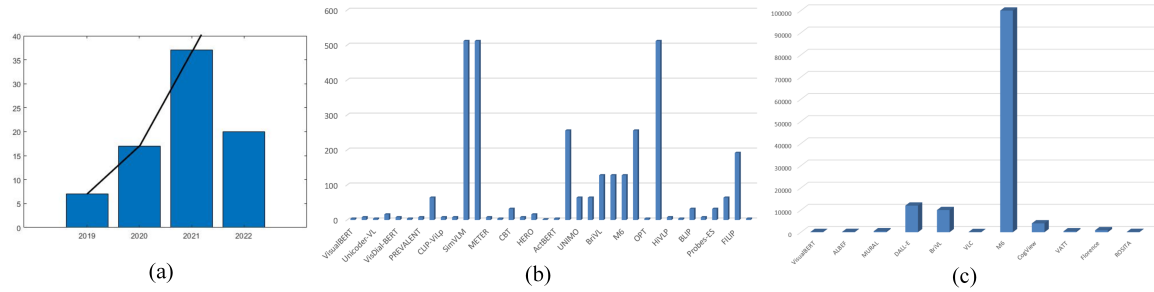


Fig. 11 (a). Number of MM-PTMs papers published from year 2019 to 2022; (b). Number of GPUs used for pre-training of selected models; (c). Parameters of selected MM-PTMs.

huge demand of computing power for pre-trained big MM-PTMs.

Based on Fig. 11 (c), it is also easy to find that many large-scale MM-PTMs are still with limited parameters, but some of them indeed reached new heights. For example, the DALLE-E [164] (12000 MB), BriVL [106] (10000 MB), M6 [103] (100000 MB), and CogView [166] (4000 MB). The reasons for this phenomenon may be as follows: 1). Many MM-PTMs are trained on several public datasets. The scale of parameters is greatly improved compared to traditional models, but not by a shocking amount. 2). The development of big models is also limited by the need for large-scale computing power, and only a few giant companies or research institutes have such computing power platforms.

5.2 Performance on Representative Downstream Tasks

Here, we report the experimental results of zero-shot image retrieval, image captioning, and visual question answering. From Fig. 12 (a), we can find that the performance of different MM-PTMs have a big difference on the zero-shot image retrieval task. The blue and red vertical bar denotes the results of Rank-1 and Rank-5, respectively. Some models achieve high performance on this task which demonstrates the effectiveness of large-scale pre-training. For example, the ALBEF [130] and METER [157] achieves $82.80, 96.30$ and $79.60, 94.96$ on both evaluation metric.

For the image captioning task, we can find that the compared models achieved close performance on the COCO dataset according to Fig. 12 (b). Specifically, OSCAR [17] obtains 41.7, 30.6, 140, 24.5; VinVL attains [113] 41, 31.1, 140.9, 25.2; SimVLM achieves [116] 40.6, 33.7, 143.3, 25.4, respectively. These results are significantly better than traditional image captioning models pre-trained in a supervised manner through ImageNet [2] classification task. Similar results can also be concluded from Fig. 12 (c).

6 Research Directions

Although the multi-modal pre-trained big models have obtained huge development, however, it is still a young research direction. Many problems and opportunities are still waiting for researchers to solve. In this section, we summarize several research points which are worthy to be tried.

• **Pre-training on More Modalities**: Existing large-scale PTMs are usually pre-trained on two modalities, e.g., the vision and language. The missing of large amount aligned multi-modal data may be a key reason. As an old saying goes, "Sharpening your axe will not delay your job of chopping wood". The acquirement of real multi-modal data is the most important thing for large-scale pre-training, as shown in Fig. 13, such as visual image, text, audio, radar, event streams, depth image, thermal image, etc. To the best of our knowledge, no imaging device can capture so many modalities at the same time. Therefore, the manufacture of multi-modal imaging equipment can be a very significant thing. The pre-trained big model based on these data may have a wider potential for applications.

• **Incremental Learning based Pre-training**: Currently, existing pre-trained big methods are used for downstream tasks through feature finetuning or prompt learning [266]. This standard deep learning procedure works well in a

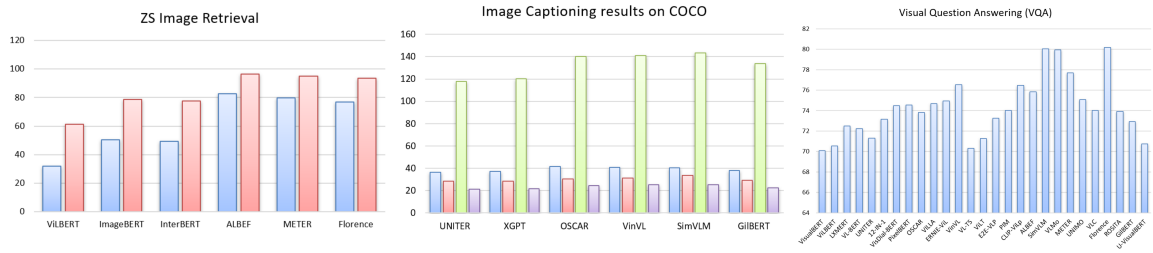


Fig. 12 Experimental results of selected MM-PTMs on zero-shot image retrieval (Rank-1, Rank-5), image captioning (BLEU, METEOR, CIDEr, SPICE), and visual question answering (Test-std).

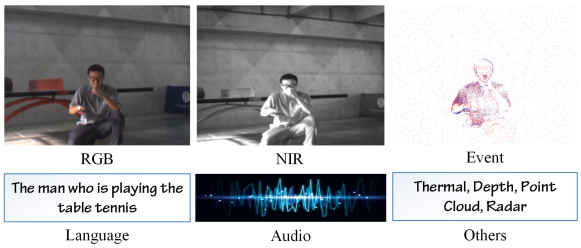


Fig. 13 Representative samples of mainstream modalities frequently used.

short time, but pre-training is an expensive process. Specifically, the collection and cleaning of data, the electric charge used for pre-training, and the hardware device all cost a huge amount of human and material resources. When we gathered another group of data, the pre-training on the mixed data are expensive, redundant, and not environmentally friendly. However, seldom of them consider incremental learning for big models, and it is still unclear if the incremental learning algorithms developed for traditional deep learning work well for big models.

In addition to the aforementioned data incremental learning, there are still many aspects that can be exploited for multi-modal pre-trained big modals. For example, the class (or category) incremental learning is a classical machine learning problem. Another interesting problem is modality-incremental learning, in another word, how to introduce and absorb the new modality into the already pre-trained multi-modal model. Because the new sensors (modalities) will appear at some indefinite time in the future, the designed multi-modal big models should be flexible enough to handle this situation.

• **Knowledge Enhanced Multi-Modal Pre-training**: Based on aforementioned reviews on MM-PTMs, we can find that the study of knowledge-assisted pre-training is still in the starting stage. Current works simply adopt external knowledge-graph or knowledge base in the pre-training phase, but they are usually single-modal, independent of multi-modal data, and limited to improving the understanding of data for models. Although commonsense knowledge is more ubiquitous, it is also abstract and introduces ambiguities, leading to challenges when applying to specific data. Therefore, we believe that further explorations on knowledge enhanced multi-modal pre-training are worth investigating. First, specified knowledge for multi-modal data is demanded to collect or extract through self-supervised learning. Second, more general knowledge fusion methods designed for multi-modal data are needed, beyond the limitations of vision and language modalities. Third, knowledge evaluation tasks specific for pre-training are required to inspect the enhancement of knowledge at this early stage, because pre-training is the first phase of the entire training procedure while downstream tasks are to be determined.

• **Fine-grained Multi-Modal Pre-training**: Most existing MM-PTMs are pre-trained from a global-view, for example, the researchers adopt the matching between the whole image and language as a supervised signal for the pre-training. The representative works are CLIP [77], ALIGN [21], etc. Note that, the fine-grained local information mining or instance-level pre-training may further improve the overall performance of multi-modal pre-training. Some researchers have exploited the possibilities of fine-grained pre-training strategies [98]. We hope more researchers can focus on this direction to further boost the final results.

• **Multi-Modal Pre-trained Model based Prompt Learning**: Current pre-trained big models are usually used in a "pretrain-finetuning"

way, specifically, the users need to initialize their model using pre-trained weights, then, finetune on downstream tasks. Although it works well in many tasks, however, the finetune maybe not be the most direct way. Because current multi-modal big models are pre-trained via modality matching, masked token prediction, and the downstream tasks are usually classification and regression tasks. Therefore, it exists a gap between multi-modal pre-training and finetuning. Recently, a new framework (termed prompt learning) is developed for big model based downstream tasks, which slickly transforms the setting of downstream tasks to make them consistent with pre-training [266]. Many works have demonstrated its effectiveness [76, 135, 261, 264, 265] in CV and NLP tasks. The research in this direction is also interesting and has great potential.

• **Migration of techniques developed for small-scale models**: The small-scale multi-modal models have been exploited for many years, and many representative models are proposed for deep multi-modal tasks [267–269]. Among these works, diffusion, cross-attention, and dynamic neural networks are useful for specific multi-modal tasks. Part of these techniques is exploited in VL-PTMs, such as the cross-attention based ViLBERT [140]. There are still many algorithms or tricks that have not yet been explored on large model tasks. We believe the transfer from small-scale to large-scale PTMs is worthy to be studied.

• **Coupling and decoupling problems in cross-modal pre-training models**: The coupling involves establishing the correlation between different modalities and the "cross" can be only realized through such correlation. The decoupling can further expand the modality dynamically. It is worth studying how to give feasible solutions to the two problems from the aspect of framework design.

7 Conclusion

We give a comprehensive review of large-scale Multi-Modal Pre-Trained Models (MM-PTMs) in this paper. Firstly, we introduce the background of MM-PTMs, with a focus on conventional deep learning, and pre-training in NLP, CV, and speech. Then, the task definition, key challenges, and benefits of MM-PTMs are discussed. After that, we dive into the reviews of MM-PTMs and discuss the pre-training data, objectives, networks, knowledge enhanced pre-training, etc. We review the downstream tasks including generative, classification, and regression tasks, and also give an overview of model parameters of MM-PTMs and hardware for the pre-training. Experimental results of several representative tasks are also discussed and visualized. Finally, we point out some research directions that are worth to be focused on. We summarize this paper and hope our survey can provide some useful insights for the MM-PTMs.

Acknowledgement

This work is supported by Key-Area Research and Development Program of Guangdong Province (No. 2021B0101400002), Peng Cheng Laboratory Key Research Project (No. PCL2021A07), Multi-source Cross-platform Video Analysis and Understanding for Intelligent Perception in Smart City (No. U20B2052), National Natural Science Foundation of China (No. 61872256, 62102205).

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

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

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[7] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[8] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. *Advances in neural information processing systems*, 28, 2015.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[10] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[11] Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, and Ming Zhou. Xgpt: Cross-modal generative pre-training for image captioning. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 786–797. Springer, 2021.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[13] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

[14] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[17] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[18] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[19] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive

language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

[20] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[22] Jing Liu, Xinxin Zhu, Fei Liu, Longteng Guo, Zijia Zhao, Mingzhen Sun, Weining Wang, Hanqing Lu, Shiyu Zhou, Jiajun Zhang, et al. Opt: Omni-perception pre-trainer for cross-modal understanding and generation. *arXiv preprint arXiv:2107.00249*, 2021.

[23] De Cheng, Jingyu Zhou, Nannan Wang, and Xinbo Gao. Hybrid dynamic contrast and probability distillation for unsupervised person re-id. *IEEE Transactions on Image Processing*, 31:3334–3346, 2022.

[24] Feilong Chen, Duzhen Zhang, Minglun Han, Xiuyi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *arXiv preprint arXiv:2202.09061*, 2022.

[25] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022.

[26] Munazza Zaib, Quan Z Sheng, and Wei Emma Zhang. A short survey of pre-trained language models for conversational ai-a new age in nlp. In *Proceedings of the Australasian Computer Science Week Multiconference*, pages 1–4, 2020.

[27] Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. A survey of controllable text generation using transformer-based pre-trained language models. *arXiv preprint arXiv:2201.05337*, 2022.

[28] Jian Yang, Gang Xiao, Yulong Shen, Wei Jiang, Xinyu Hu, Ying Zhang, and Jinghui Peng. A survey of knowledge enhanced pre-trained models. *arXiv preprint arXiv:2110.00269*, 2021.

[29] Da Yin, Li Dong, Hao Cheng, Xiaodong Liu, Kai-Wei Chang, Furu Wei, and Jianfeng Gao. A survey of knowledge-intensive nlp with pre-trained language models. *arXiv preprint arXiv:2202.08772*, 2022.

[30] Prajjwal Bhargava and Vincent Ng. Commonsense knowledge reasoning and generation with pre-trained language models: A survey. *arXiv preprint arXiv:2201.12438*, 2022.

[31] Qi Liu, Matt J Kusner, and Phil Blunsom. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*, 2020.

[32] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021.

[33] Benyou Wang, Qianqian Xie, Jiahuan Pei, Prayag Tiwari, Zhao Li, et al. Pre-trained language models in biomedical domain: A systematic survey. *arXiv preprint arXiv:2110.05006*, 2021.

[34] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897, 2020.

[35] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. Pre-trained models: Past, present and

future. *AI Open*, 2:225–250, 2021.

[36] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022.

[37] Ludan Ruan and Qin Jin. Survey: Transformer based video-language pre-training. *AI Open*, 2022.

[38] Feng Li, Hao Zhang, Yi-Fan Zhang, Shilong Liu, Jian Guo, Lionel M Ni, PengChuan Zhang, and Lei Zhang. Vision-language intelligence: Tasks, representation learning, and large models. *arXiv preprint arXiv:2203.01922*, 2022.

[39] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[40] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 2021.

[41] Yang Liu, Yao Zhang, Yixin Wang, Feng Hou, Jin Yuan, Jiang Tian, Yang Zhang, Zhongchao Shi, Jianping Fan, and Zhiqiang He. A survey of visual transformers. *arXiv preprint arXiv:2111.06091*, 2021.

[42] Javier Selva, Anders S Johansen, Sergio Escalera, Kamal Nasrollahi, Thomas B Moeslund, and Albert Clapés. Video transformers: A survey. *arXiv preprint arXiv:2201.05991*, 2022.

[43] Shangwei Guo, Chunlong Xie, Jiwei Li, Lingjuan Lyu, and Tianwei Zhang. Threats to pre-trained language models: Survey and taxonomy. *arXiv preprint arXiv:2202.06862*, 2022.

[44] Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184, 2021.

[45] Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. *arXiv preprint arXiv:2110.08527*, 2021.

[46] Rohit Kumar Kaliyar. A multi-layer bidirectional transformer encoder for pre-trained word embedding: A survey of bert. In *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 336–340. IEEE, 2020.

[47] Jiajia Peng and Kaixu Han. Survey of pre-trained models for natural language processing. In *2021 International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB)*, pages 277–280. IEEE, 2021.

[48] Sha Yuan, Hanyu Zhao, Shuai Zhao, Jiahong Leng, Yangxiao Liang, Xiaozhi Wang, Jifan Yu, Xin Lv, Zhou Shao, Jiaao He, et al. A roadmap for big model. *arXiv preprint arXiv:2203.14101*, 2022.

[49] Soyeon Caren Han Siqu Long, Feiqi Cao and Haiqing Yang. Vision-and-language pretrained models: A survey. In *IJCAI*, 2022.

[50] Xu Peng, Zhu Xiatian, and A. Clifton David. Multimodal learning with transformers: A survey. *arXiv preprint arXiv:2206.06488*, 2022.

[51] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[52] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[53] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897, 2020.

[54] Munazza Zaib, Quan Z Sheng, and Wei Emma Zhang. A short survey of pre-trained language models for conversational ai-a new age in nlp. In *Proceedings of the Australasian Computer Science Week Multiconference*, pages 1–4, 2020.

[55] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*, 2021.

[56] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021.

[57] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[58] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[59] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[60] Corby Rosset. Turing-nlg: A 17-billion-parameter language model by microsoft. *Microsoft Blog*, 1(2), 2020.

[61] Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, et al. Pangu-*alpha*: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*, 2021.

[62] Junqiu Wei, Xiaozhe Ren, Xiaoguang Li, Wenyong Huang, Yi Liao, Yasheng Wang, Jiashu Lin, Xin Jiang, Xiao Chen, and Qun Liu. Nezha: Neural contextualized representation for chinese language understanding. *arXiv preprint arXiv:1909.00204*, 2019.

[63] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020.

[64] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[65] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[66] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[67] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen

Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021.

[68] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[69] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[70] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021.

[71] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[72] Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. Effectiveness of self-supervised pre-training for speech recognition. *arXiv preprint arXiv:1911.03912*, 2019.

[73] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[74] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[75] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *arXiv preprint arXiv:2108.06209*, 2021.

[76] Peipei Zhu, Xiao Wang, Lin Zhu, Zhenglong Sun, Weishi Zheng, Yaowei Wang, and Changwen Chen. Prompt-based learning for unpaired image captioning. *arXiv preprint arXiv:2205.13125*, 2022.

[77] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[78] Yinghui Xing, Qirui Wu, De Cheng, Shizhou Zhang, Guoqiang Liang, and Yanning Zhang. Class-aware visual prompt tuning for vision-language pre-trained model. *arXiv preprint arXiv:2208.08340*, 2022.

[79] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[80] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[81] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[82] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia

Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[83] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[84] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018.

[85] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[86] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[87] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.

[88] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[89] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[90] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, 2018.

[91] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019.

[92] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021.

[93] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[94] Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*, number FZJ-2022-00923. Jülich Supercomputing Center, 2021.

[95] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for

the people. In *NeurIPS Datasets and Benchmarks*, 2021.

[96] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Hang Xu, Xiaodan Liang, Wei Zhang, Xin Jiang, and Chunjing Xu. Wukong: 100 million large-scale chinese cross-modal pre-training dataset and a foundation framework, 2022.

[97] Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for ms-coco. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2855–2870, 2021.

[98] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11782–11791, 2021.

[99] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449, 2021.

[100] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[101] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021.

[102] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[103] Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. M6: A chinese multimodal pretrainer. *arXiv preprint arXiv:2103.00823*, 2021.

[104] Xiao Dong, Xunlin Zhan, Yangxin Wu, Yunchao Wei, Xiaoyong Wei, Minlong Lu, and Xiaodan Liang. M5product: A multi-modal pretraining benchmark for e-commercial product downstream tasks. *arXiv preprint arXiv:2109.04275*, 2021.

[105] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European Conference on Computer Vision*, pages 647–664. Springer, 2020.

[106] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*, 2021.

[107] Leng Jiahong Xue Zhao Zhao Hanyu Sha Yuan, Zhao Shuai and Tang Jie. Wudaomm: A large-scale multi-modal dataset for pre-training models. *arXiv preprint arXiv:2203.11480*, 2022.

[108] Delong Chen, Fan Liu, Xiaoyu Du, Ruizhuo Gao, and Feng Xu. Mep-3m: A large-scale multi-modal e-commerce products dataset.

[109] Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. Wenlan 2.0: Make ai imagine via a multimodal foundation model. *arXiv preprint arXiv:2110.14378*, 2021.

[110] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[111] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[112] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[113] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[114] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020.

[115] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. Interbert: Vision-and-language interaction for multi-modal pretraining. *arXiv preprint arXiv:2003.13198*, 2020.

[116] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pre-training with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[117] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019.

[118] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[119] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems*, 32, 2019.

[120] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[121] Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. A fast proximal point method for computing exact wasserstein distance. In *Uncertainty in artificial intelligence*, pages 433–453. PMLR, 2020.

[122] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13137–13146, 2020.

[123] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang.

Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216, 2021.

[124] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. Kaleido-bert: Vision-language pre-training on fashion domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12647–12657, 2021.

[125] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. E2e-vlp: End-to-end vision-language pre-training enhanced by visual learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 503–513, 2021.

[126] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, 2020.

[127] Yan Ling, Rui Xia, et al. Vision-language pre-training for multimodal aspect-based sentiment analysis. *arXiv preprint arXiv:2204.07955*, 2022.

[128] Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. Rosita: Enhancing vision-and-language semantic alignments via cross- and intra-modal knowledge integration. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 797–806, 2021.

[129] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, pages 1–38, 2022.

[130] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021.

[131] Ziyi Yang, Yuwei Fang, Chenguang Zhu, Reid Pryzant, Dongdong Chen, Yu Shi, Yichong Xu, Yao Qian, Mei Gao, Yi-Ling Chen, et al. i-code: An integrative and composable multimodal learning framework. *arXiv preprint arXiv:2205.01818*, 2022.

[132] Wei Suo, Mengyang Sun, Peng Wang, and Qi Wu. Proposal-free one-stage referring expression via grid-word cross-attention. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 1032–1038. ijcai.org, 2021.

[133] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020.

[134] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[135] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. *arXiv preprint arXiv:2201.05078*, 2022.

[136] Yufeng Cui, Lichen Zhao, Feng Liang, Yangguang Li, and Jing Shao. Democratizing contrastive language-image pre-training: A clip benchmark of data, model, and supervision. *arXiv preprint arXiv:2203.05796*, 2022.

[137] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.

[138] Chen Delong, Wu Zhao, Liu Fan, Yang Zaiquan, Huang Yixiang, Bao Yiping, and Zhou Erjin. Prototypical contrastive language image pretraining. *arXiv preprint arXiv:2206.10996*, 2022.

[139] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[140] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[141] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2131–2140, 2019.

[142] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019.

[143] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020.

[144] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

[145] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *European Conference on Computer Vision*, pages 336–352. Springer, 2020.

[146] Gao Yuting, Liu Jinfeng, Xu Zihan, Zhang Jun, Li Ke, and Shen Chunhua. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. In *arXiv:2204.14095*, 2022.

[147] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2251–2260, 2020.

[148] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020.

[149] Dandan Song, Siyi Ma, Zhanchen Sun, Sicheng Yang, and Lejian Liao. Kvl-bert: Knowledge enhanced visual-and-linguistic bert for visual commonsense reasoning. *Knowledge-Based Systems*, 230:107408, 2021.

[150] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.

[151] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[152] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[153] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021.

[154] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training. *Advances in Neural Information Processing Systems*, 34, 2021.

[155] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. Mural: multimodal, multitask retrieval across languages. *arXiv preprint arXiv:2109.05125*, 2021.

[156] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021.

[157] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Zicheng Liu, Michael Zeng, et al. An empirical study of training end-to-end vision-and-language transformers. *arXiv preprint arXiv:2111.02387*, 2021.

[158] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019.

[159] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.

[160] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

[161] Aisha Urooj, Amir Mazaheri, Mubarak Shah, et al. Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4648–4660, 2020.

[162] Rui Yan, Mike Zheng Shou, Yixiao Ge, Alex Jinpeng Wang, Xudong Lin, Guanyu Cai, and Jinhui Tang. Video-text pre-training with learned regions. *arXiv preprint arXiv:2112.01194*, 2021.

[163] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

[164] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[165] Alex Hauptmann Yonatan Bisk Jianfeng Gao Liangke Gui, Qiuyuan Huang. Training vision-language transformers from captions alone. *arXiv preprint arXiv:2205.09256*, 2022.

[166] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers.

Advances in Neural Information Processing Systems, 34, 2021.

[167] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34, 2021.

[168] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[169] Mickael Coustaty Marçal Rusiñol Oriol Ramos Terrades Souhail Bakkali, Zuheng Ming. Hivlp: Hierarchical vision-language pre-training for fast image-text retrieval. *arXiv preprint arXiv:2205.12029*, 2022.

[170] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. *arXiv preprint arXiv:2203.05175*, 2022.

[171] Weixiang Hong, Kaixiang Ji, Jiajia Liu, Jian Wang, Jingdong Chen, and Wei Chu. Gilbert: Generative vision-language pre-training for image-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1379–1388, 2021.

[172] Yuqi Huo Yizhao Gao Zhiwu Lu Ji-Rong Wen Haoyu Lu, Nanyi Fei. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *arXiv:2204.07441*, 2022.

[173] Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. Unsupervised vision-and-language pre-training without parallel images and captions. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5339–5350, 2021.

[174] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[175] Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3977–3986, 2021.

[176] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[177] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. N\" uwa: Visual synthesis pre-training for neural visual world creation. *arXiv preprint arXiv:2111.12417*, 2021.

[178] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. *arXiv preprint arXiv:2202.10401*, 2022.

[179] Minlong Lu Wei, Yaowei Wang, and Xiaodan Liang. M5product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining.

[180] Bin Yan and Mingtao Pei. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. 2022.

[181] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai,

Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. *arXiv preprint arXiv:2112.09106*, 2021.

[182] Xiwen Liang, Fengda Zhu, Lingling Li, Hang Xu, and Xiaodan Liang. Visual-language navigation pretraining via prompt-based environmental self-exploration. *arXiv preprint arXiv:2203.04006*, 2022.

[183] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022.

[184] Xie Chunyu, Cai Heng, Song Jianfei, Li Jincheng, Kong Fanjing, Wu Xiaoyu, Morimitsu Henrique, Yao Lin, Wang Dexin, Leng Dawei, Ji Xiangyang, and Deng Yafeng. Zero and r2d2: A large-scale chinese cross-modal benchmark and a vision-language framework. *arXiv preprint arXiv:2205.03860*, 2022.

[185] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.

[186] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

[187] Chenliang Li, Ming Yan, Haiyang Xu, Fuli Luo, Wei Wang, Bin Bi, and Songfang Huang. Semvlp: Vision-language pre-training by aligning semantics at multiple levels. *arXiv preprint arXiv:2103.07829*, 2021.

[188] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[189] Jiaxin Shi Duzhen Zhang Jianlong Chang Feilong Chen, Xiuyi Chen and Qi Tian. Hivlp: Hierarchical vision-language pre-training for fast image-text retrieval. *arXiv preprint arXiv:2205.12105*, 2022.

[190] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.

[191] Hangbo Bao, Wenhui Wang, Li Dong, and Furu Wei. Vl-beit: Generative vision-language pretraining. *arXiv preprint arXiv:2206.01127*, 2022.

[192] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. *arXiv preprint arXiv:2201.08264*, 2022.

[193] Fan Zhihao, Wei Zhongyu, Chen Jingjing, Wang Siyuan, Li Zejun, Xu Jiarong, and Huang Xuanjing. A unified continuous learning framework for multi-modal knowledge discovery and pre-training. *arXiv preprint arXiv:2206.05555*, 2022.

[194] Zhang Haotian, Zhang Pengchuan, Hu Xiaowei, Chen Yen-Chun, Harold Li Liunian, Dai Xiyang, Wang Lijuan, Yuan Lu, Hwang Jenq-Neng, and Gao Jianfeng. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022.

[195] Mustafa Basil, Riquelme Carlos, Puigcerver Joan, Jenatton Rodolphe, and Houlsby Neil. Multimodal contrastive learning with limoe: the language-image mixture of experts. *arXiv preprint arXiv:2206.02770*, 2022.

[196] Wang Teng, Jiang Wenhao, Lu Zhichao, Zheng Feng, Cheng Ran, Yin Chengguo, and Ping Luo. Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix. *arXiv preprint arXiv:2206.08919*, 2022.

[197] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[198] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1), Jun. 2014.

[199] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 687–696, Beijing, China, July 2015. Association for Computational Linguistics.

[200] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2181–2187. AAAI Press, 2015.

[201] Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. Knowledge graph completion with adaptive sparse transfer matrix. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Feb. 2016.

[202] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 809–816, Madison, WI, USA, 2011. Omnipress.

[203] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[204] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases, 2014.

[205] Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. A semantic matching energy function for learning with multi-relational data. *Machine Learning*, 94(2):233–259, Feb 2014.

[206] Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. Holographic embeddings of knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar. 2016.

[207] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann Lecun. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations (ICLR2014), CBLS, April 2014*, 2014.

[208] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[209] Thomas N. Kipf and Max Welling. Variational graph auto-encoders, 2016.

[210] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[211] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2017.

[212] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, *The Semantic Web*, pages 593–607, Cham, 2018. Springer International Publishing.

[213] Chao Shang, Yun Tang, Jing Huang, Jinbo Bi, Xiaodong He, and Bowen Zhou. End-to-end structure-aware convolutional networks for knowledge base completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3060–3067, Jul. 2019.

[214] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[215] Deepak Nathani, Jatin Chauhan, Charu Sharma, and Manohar Kaul. Learning attention-based embeddings for relation prediction in knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019.

[216] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. Composition-based multi-relational graph convolutional networks. In *International Conference on Learning Representations*, 2020.

[217] Yanzeng Li, Bowen Yu, Xue Mengge, and Tingwen Liu. Enhancing pre-trained Chinese character representation with word-aligned attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3442–3448, Online, July 2020. Association for Computational Linguistics.

[218] Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. SentiLARE: Sentiment-aware language representation learning with linguistic knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6975–6988, Online, November 2020. Association for Computational Linguistics.

[219] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online, November 2020. Association for Computational Linguistics.

[220] Devendra Sachan, Yuhao Zhang, Peng Qi, and William L. Hamilton. Do syntax trees help pre-trained transformers extract information? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online, April 2021. Association for Computational Linguistics.

[221] Junru Zhou, Zhuosheng Zhang, Hai Zhao, and Shuailiang Zhang. LIMIT-BERT : Linguistics informed multi-task BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4450–4461, Online, November 2020. Association for Computational Linguistics.

[222] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: Enhanced language representation with informative entities. In *Proceedings of ACL 2019*, 2019.

[223] Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, Hong Kong, China, November 2019. Association for Computational Linguistics.

[224] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Explicit knowledge-based reasoning for

visual question answering. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1290–1296, 2017.

[225] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Fvqa: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2413–2427, 2018.

[226] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 48–64, Cham, 2014. Springer International Publishing.

[227] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

[228] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[229] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[230] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[231] Zhaochen Guo and Denilson Barbosa. Robust entity linking via random walks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, page 499–508, New York, NY, USA, 2014. Association for Computing Machinery.

[232] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[233] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. Abductive commonsense reasoning. In *International Conference on Learning Representations*, 2020.

[234] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. CommonGen: A constrained text generation challenge for generative commonsense reasoning.

In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online, November 2020. Association for Computational Linguistics.

[235] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[236] Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439, Apr. 2020.

[237] Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China, November 2019. Association for Computational Linguistics.

[238] Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. Temporal reasoning on implicit events from distant supervision. In *NAACL*, 2021.

[239] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957, 2019.

[240] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017.

[241] Pengcheng Yang, Boxing Chen, Pei Zhang, and Xu Sun. Visual agreement regularized training for multi-modal machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9418–9425, 2020.

[242] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[243] Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10900–10910, 2020.

[244] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, 2017.

[245] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

[246] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer, 2005.

[247] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.

[248] Xiao Wang, Shaofei Zheng, Rui Yang, Aihua Zheng, Zhe Chen, Jin Tang, and Bin Luo. Pedestrian attribute recognition: A survey. *Pattern Recognition*, 121:108220, 2022.

[249] Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Contextual inter-modal attention for multi-modal sentiment analysis. In *proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3454–3466, 2018.

[250] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1970–1979, 2017.

[251] Wei Chen, Yang Liu, Weiping Wang, Erwin M Bakker, TK Georgiou, Paul Fieguth, Li Liu, and MSK Lew. Deep image retrieval: A survey. *ArXiv*, 2021.

[252] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Eric Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. *arXiv preprint arXiv:2203.12667*, 2022.

[253] Sang-Min Park and Young-Gab Kim. Visual language navigation: a survey and open challenges. *Artificial Intelligence Review*, pages 1–63, 2022.

[254] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4158–4166, 2018.

[255] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4145–4154, 2019.

[256] Xinpeng Ding, Nannan Wang, Shiwei Zhang, Ziyuan Huang, Xiaomeng Li, Mingqian Tang, Tongliang Liu, and Xinbo Gao. Exploring language hierarchy for video grounding. *IEEE Transactions on Image Processing*, 31:4693–4706, 2022.

[257] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[258] Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13763–13773, 2021.

[259] Xiao Wang, Chenglong Li, Rui Yang, Tianzhu Zhang, Jin Tang, and Bin Luo. Describe and attend to track: Learning natural language guided structural representation and visual attention for object tracking. *arXiv preprint arXiv:1811.10014*, 2018.

[260] Qi Feng, Vitaly Ablavsky, Qinxun Bai, and Stan Sclaroff. Siamese natural language tracker: Tracking by natural language descriptions with siamese trackers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5851–5860, 2021.

[261] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021.

[262] Xuehai He, Diji Yang, Weixi Feng, Tsu-Jui Fu, Arjun Akula, Varun Jampani, Pradyumna Narayana, Sugato Basu, William Yang Wang, and Xin Eric Wang. Cpl: Counterfactual prompt learning for vision and language models. *arXiv preprint arXiv:2210.10362*, 2022.

[263] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022.

[264] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[265] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.

[266] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021.

[267] Qingzheng Wang, Shuai Li, Hong Qin, and Aimin Hao. Robust multi-modal medical image fusion via anisotropic heat diffusion guided low-rank structural analysis. *Information fusion*, 26:103–121, 2015.

[268] Xiao Wang, Xiujun Shu, Shilliang Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Mfgnet: Dynamic modality-aware filter generation for rgb-t tracking. *IEEE Transactions on Multimedia*, 2022.

[269] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.