Vision-and-Language Pretrained Models: A Survey

Siqu Long[1] , **Feiqi Cao**[1] , **Soyeon Caren Han**[1] and **Haiqin Yang**[2*]
[1]School of Computer Science, The University of Sydney, Australia
[2]International Digital Economy Academy (IDEA), China
{slon6753, fcao0492}@uni.sydney.edu.au, caren.han@sydney.edu.au, hqyang@ieee.org

Abstract

Pretrained models have produced great success in both Computer Vision (CV) and Natural Language Processing (NLP). This progress leads to learning joint representations of vision and language pretraining by feeding visual and linguistic contents into a multi-layer transformer, Visual-Language Pretrained Models (VLPMs). In this paper, we present an overview of the major advances achieved in VLPMs for producing joint representations of vision and language. As the preliminaries, we briefly describe the general task definition and genetic architecture of VLPMs. We first discuss the language and vision data encoding methods and then present the mainstream VLPM structure as the core content. We further summarise several essential pretraining and fine-tuning strategies. Finally, we highlight three future directions for both CV and NLP researchers to provide insightful guidance.

1 Introduction

In both Computer Vision (CV) and Natural Language Processing (NLP) communities, pretrained models have made significant progress. While CV researchers use VGG and ResNet using ImageNet to predict the categorical label of a given image, BERT [Devlin *et al.*, 2019] has been used and revolutionised many NLP tasks, such as natural language inference, and reading comprehension. Motivated by this, many cross-modal Vision-Language Pretrained Models (VLPMs) have been designed [Lu *et al.*, 2019; Su *et al.*, 2019; Chen *et al.*, 2020; Li *et al.*, 2020b]. This pretrain-then-transfer learning approach to vision-language tasks naturally follows its widespread use in both CV and NLP. It has become the de facto standard due to the ease of use and solid representational power of large, publicly available models trained on large-scaled data sources.

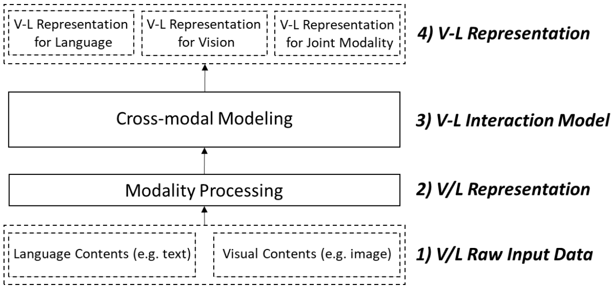
V-L Representation for Language
V-L Representation for Vision
V-L Representation for Joint Modality
4) V-L Representation
Cross-modal Modeling
3) V-L Interaction Model
Modality Processing
2) V/L Representation
Language Contents (e.g. text)
Visual Contents (e.g. image)
1) V/L Raw Input Data

Figure 1: General architecture of VLPMs

In this paper, we present an overview of the rise and major advances achieved in the topic of VLPMs. Figure 1 illustrates a generic architecture of VLPMs. It involves the design of four main components: **1) V/L (Vision and Language) Raw Input Data** defines the representative *raw data streams* from language and visual contents respectively, such as a single or multiple sentence(s) and one or a set of image(s). **2) V/L Representation** processes the raw data input into the desired format of *modality representations* that can be used for **3) V-L (Vision-Language) Interaction Model**, which then enforces the cross-modal modeling between the two modalities. For instance, a common design is that the textual sentence is first tokenized and converted into the Bert-formatted input embedding while the image is processed into a set of spatial-aware RoI (Region of Interest) features. Those two modality representations are then concatenated and fed into the transformer encoder layers in which the cross-modal interaction is modeled via the multi-head self-attention mechanism. **4) V-L Representation** defines the possible cross-modal representations, which can be a *V-L representation* for the single modality (i.e., Language or Vision) and/or the *V-L representation* of joint modalities (i.e., Language and Vision). With the well-designed task supervision and learning guidelines from the pretraining, the *V-L representation* finally learns to represent the generic cross-modal semantics, which would be transferred to help with the downstream V-L tasks via fine-tuning. This generic architecture applies to most of the existing VLPMs. The designs are various for each component, their pretraining strategies and transfer applications.

Existing surveys in this area have only partially reviewed some related tasks [Mogadala *et al.*, 2021] or focused mainly on systematical analysis [Bugliarello *et al.*, 2021]. To the best of our knowledge, this is the first work that presents a comprehensive review of VLPMs. Our paper aims to provide both CV and NLP researchers insightful guidance for visual and language cross-modal learning via pretraining.

[*]Corresponding author: hqyang@ieee.org. Work done when Siqu and Feiqi were interned at IDEA.

2 Input Data Encoding

2.1 Language Encoding

Most VLPMs represent the *language input* by taking a single textual sentence, which directly aligns with the image (visual modality) since their pretraining process mainly relies on pairwise image-text corpus [Li *et al.*, 2019; Tan and Bansal, 2019; Lu *et al.*, 2019; Li *et al.*, 2020a; Chen *et al.*, 2020; Yu *et al.*, 2021; Li *et al.*, 2020b]. Some VLPMs use multiple sentences for representing the language input, especially with visual dialogue and multi-lingual settings [Wang *et al.*, 2020; Ni *et al.*, 2021; Fei *et al.*, 2021]. Some special VLPMs apply visual input in the form of text and encode it as a part of the language input [Li *et al.*, 2020b; Yang *et al.*, 2021]. For instance, OSCAR appends a set of object class tags detected from an image to the textual sentence as a language input to learn object-aligned V-L representation [Li *et al.*, 2020b]. This early-fusion strategy of vision to language (V2L) at the raw input level serves as an anchor point to ease the cross-modal alignment learning.

To process the *language input* into a suitable *language representation* for cross-modal modeling, almost all VLPMs directly adapt the Bert-formatted input representation [Devlin *et al.*, 2019], which sums up three types of learnable embeddings for each token in the textual sequence: 1) *token embedding*, 2) *position embedding*, and 3) *segment embedding*. The *token embedding* mostly follows the original Bert and encodes the textual sequence in the general form of "[CLS]$w_1 w_2 ... w_i$[SEP]" or with some minor modifications, where w_i represents the ith tokenized (sub-)word based on the WordPiece vocabulary and [CLS]/[SEP] are special tokens indicating the starting and ending of this sequence. Some models introduce additional special tokens to better align with specific data or pretraining tasks [Wang *et al.*, 2020; Zhou *et al.*, 2020; Xia *et al.*, 2021], such as the [EOT] tokens for separating each dialogue turns [Wang *et al.*, 2020]. The *position embedding* is used exactly the same as in the original Bert, but the *segment embedding* is adjusted to be *modality embedding* in order to differentiate the two modalities or to further distinguish the multiple data streams within the language modality when multiple sentences are used, e.g., the segment tokens $A/B/C$ used in VL-Bert [Su *et al.*, 2019]. As a special case, the *visual feature* (See Sec.2.2) can also be included into the language representation as the fourth type of embedding, which can be regarded as an early-fusion strategy of vision to language (V2L) at the representation level [Su *et al.*, 2019; Chiou *et al.*, 2021]. With the Bert input representation format, it allows the VLPMs to enable initialising input embedding from Bert and directly adopt transformer encoder or its variants for the further cross-modal interaction modeling with a multi-head self-attention mechanism.

Intra-modality Processing. In particular, some VLPMs apply *additional intra-modality processing* with self-attention-based *transformer* blocks to the aforementioned Bert-formatted language embeddings for further encoding the intra-modal contextual information, in order to better balance with the already high-level visual feature produced by the deep CNN-based extractor from the vision modality (see Sec. 2.2) and enable a robust single-modal representation of language [Lu *et al.*, 2019; Tan and Bansal, 2019; Li *et al.*, 2021b; Yang *et al.*, 2021; Majumdar *et al.*, 2020]. Instead, the resultant transformer output representation would be used as input for the V-L interaction model.

2.2 Vision Encoding

With the processing of language input, the *visual input* for VLPMs is normally a single image that directly aligns with the paired text input [Li *et al.*, 2019; Tan and Bansal, 2019; Lu *et al.*, 2019; Li *et al.*, 2020a; Chen *et al.*, 2020; Yu *et al.*, 2021; Li *et al.*, 2020b], or a set of image(s) that are semantically correlated with each other [Majumdar *et al.*, 2020; Hao *et al.*, 2020]. For instance, the pretraining data for Vision-Language Navigation(VLN) is formed by a textual instruction with a group of panorama images along the navigation trajectory path [Majumdar *et al.*, 2020].

Similar to the language representation, most studies encode *visual input* into the Bert-style sequential representation, which consists of the aforementioned three major embeddings. The *segment embedding* is created the same as that in language representation, but the *token embedding* and *position embedding* are modified to *visual feature* and *spatial position embedding* for capturing the visual semantics. Specifically, the *visual feature* is extracted using the CNN-based feature extractors, the most common way in the CV domain. For this, the ***granularity of representation***, i.e., the grouping of image pixels into the sequential visual tokens, decides the alignment level of cross-modal modeling in the image content: *1) RoI*-based VLPMs typically apply a pretrained Faster R-CNN object detector with a CNN backbone and extract the visual features of the detected object regions for the visual tokens [Lu *et al.*, 2019; Tan and Bansal, 2019; Wang *et al.*, 2020; Chen *et al.*, 2020; Li *et al.*, 2020a; Zhou *et al.*, 2020] This is under the assumption that most image-text pairwise data is supposed to have its text describe the salient object (regions) in the corresponding image. Comparatively, some VLPMs simply split the whole image into continuous *2) patches* [Gao *et al.*, 2020; Wang *et al.*, 2021b; Wen *et al.*, 2021; Huang *et al.*, 2021] or even more fine-grained *3) pixels* [Huang *et al.*, 2020] as visual tokens for CNN-based feature extraction and refine the visual representation end-to-end, leading to significant speed improvement. More recent studies even propose the shallow, convolution-free embedding that utilizes simple linear projection to encode visual tokens at the *patches/pixels* level as in ViT [Dosovitskiy *et al.*, 2020] for a further efficiency gain [Wang *et al.*, 2021a; Singh *et al.*, 2021; Kim *et al.*, 2021]. Besides, there are also VLPMs encoding each integral *4) image* as a visual token by taking the pooling layer result from a CNN model (e.g., EfficientNet) [Hao *et al.*, 2020].

Unlike the textual tokens with sequential positional relation, visual tokens entail ***spatial positional relation*** instead, varying based on different granularity, which can be encoded by *spatial position embedding*. RoI-based VLPMs commonly adopt the *coordinate-based* position embedding[Lu *et al.*, 2019; Li *et al.*, 2020a; Zhou *et al.*, 2020; Chen *et al.*, 2020], such as the 5-dimensional vector representing the normalized coordinates of the RoI bounding boxes and the fraction of image area. Comparatively, pixel/patch-based VLPMs represent the *pixel location* using the 2D-aware vector for the row/column number [Huang *et al.*, 2021]. PREVALENT [Hao *et al.*, 2020] for Vision-Language Navigation (VLN) is the only image-based VLPM that specifies the spatial position embedding, which uses the elevation and heading angle relative to the agent to represent the positional relations among panorama images. The *spatial position embedding* is normally combined with the *visual feature* with or without *segment embedding* via one or two FC layers followed by Layer Normalization to form the *visual representation*.

Intra-modality Processing. Since the transformer block used for additional intra-modality processing in language (See Sec.2.1) is a global operator whereas the CNN-based visual feature extractor is a local operator, it may lead to a different feature distribution between the two modalities [Wen *et al.*, 2021]. Thus, some VLPMs also apply self-attention-based *transformer* blocks to the initial Bert-style sequential visual representation to align with the language intra-modality processing [Tan and Bansal, 2019; Hao *et al.*, 2020; Sun *et al.*, 2021; Wen *et al.*, 2021]. Using transformers for intra-modality processing enables a *modality-customized* encoding with the freedom of selecting a different number of blocks for each modality. In practice, the language modality normally uses more transformer blocks than the vision modality for a better balance [Tan and Bansal, 2019; Lu *et al.*, 2019; Hao *et al.*, 2020]. Besides, several models adopt *non-Transformer* processing, e.g., AoANet [Xia *et al.*, 2021] and Visual Dictionary mapping [Huang *et al.*, 2021].

3 V-L Interaction Model (V-LIM)

There are two types of mainstream VLPM model structures: **(1) Single-stream (1-stream)** [Li *et al.*, 2019; Chen *et al.*, 2020; Li *et al.*, 2020b], which directly fuse the *initial language/visual representation* by using the joint cross-modal encoder at the initial stage, and **(2) Double-stream (2-stream)** [Tan and Bansal, 2019; Lu *et al.*, 2019; Murahari *et al.*, 2020], which separately apply the *intra-modality processing* to two modalities along with a shared cross-modal encoder. However, this way of classification is mainly based on the perspective of intra-modality data-handling. In this section, we briefly review the model structure and the major output of VLPMs by focusing on their *V-L interaction models (V-LIM)* instead, which can be: **1) Self-attention-based**, **2) Co-attention-based** or **3) VSE-based V-LIM**.

1) Self-attention-based V-LIM. Most single-stream [Li *et al.*, 2019; Li *et al.*, 2020a; Su *et al.*, 2019; Huang *et al.*, 2020; Gao *et al.*, 2020; Chen *et al.*, 2020; Li *et al.*, 2020b] and some of the double-stream VLPMs [Huang *et al.*, 2021] are considered as a *self-attention-based V-LIMs* since they directly apply single-stream self-attention module to the modality representations for cross-modal modeling. They simply concatenate the language and visual representation (as introduced in Sec.2.1 and Sec.2.2) to produce an integrated sequence representation so that a self-attention stream in the following transformer blocks can enforce the dense interaction modeling for both intra-modal (V-V&L-L) and cross-modal (V-L). Following the transformer, the output representation of the global-level special token, such as [CLS], in the multi-modal sequence would be taken as the holistic *joint V-L representation* for both pretraining and transfer learning. The output representation in the language or vision sequence could represent the contextualized V-L semantics and thus can be used as the *V/L representation for Language or Vision*.

2) Co-attention-based V-LIM. Comparatively, VLPMs with *co-attention-based V-LIM* decouple the intra- and cross-modal modeling processes. They keep two separate streams of transformer blocks for each modality that are entangled only at the specific *cross-attention sub-layer(s)*. These sub-layers enforce exchange of the *key* and *value* in self-attention module between the two modality streams. In this way, they limit the cross-modal modeling only to those co-attention sub-layers while leaving the rest of the sub-layers independent to focus on intra-modality processing and modeling. This single-modality focus makes them align well with the definition of double-stream VLPMs [Tan and Bansal, 2019; Lu *et al.*, 2019; Murahari *et al.*, 2020; Majumdar *et al.*, 2020; Hao *et al.*, 2020; Yu *et al.*, 2021]. As a result, the *V/L representation for Language or Vision* can be taken from the output representation of the two separate modality streams whereas the *V-L representation for joint modality* is derived by taking the global token representation from either modality sequence (or their aggregation).

Encoder-decoder VLPMs. There are some VLPMs that apply a transformer-based encoder-decoder structure to empower the V-L generation ability. They can be categorized into either single-stream [Zhou *et al.*, 2020; Wang *et al.*, 2020; Wang *et al.*, 2021b; Li *et al.*, 2021a] or double-stream [Xia *et al.*, 2021; Li *et al.*, 2021b]. Regarding the V-LIM, few of them adopt *unified encoder-decoder* structure with *self-attention-based V-LIM* [Zhou *et al.*, 2020; Wang *et al.*, 2020; Li *et al.*, 2021a]. Others try to emphasize the different peculiarities of the understanding/generation disciplines and instead use the conventional *decoupled encoder-decoder* structure [Wang *et al.*, 2021b; Xia *et al.*, 2021; Li *et al.*, 2021b]. Their V-LIM can be either *self-* or *co-attention-based* or even their combination, depending on how the attention modules in encoder/decoder are utilized for handling the multi-modalities. In general, these encoder-decoder VLPMs derive the *V-L* representation via the cross-modal encoder in a similar way to those self-attention-based VLPMs.

3) VSE-based V-LIM. More recently, there is another emerging mainstream of VLPMs that simply utilizes the dual-encoder structured model with Visual-Semantic-Embedding(VSE)-based cross-modal contrastive learning [Radford *et al.*, 2021; Jia *et al.*, 2021; Sun *et al.*, 2021; Wen *et al.*, 2021]. They tend to have a special focus on large-scale cross-modal retrieval with high demand for efficiency. They utilize the intra-modality processing to derive the vision and language representation (as described in Sec.2.1 and Sec.2.2) between which the similarity-based cross-modal alignment would be modeled in the shared VSE space at the global level. This dual-encoder structure eradicates the fusion-style attention-based V-LIMs that are computation-costly and time-consuming. At the same time, it enables independent V/L representation encoding of both modalities, making the pre-computation possible for more efficient retrieval. The resultant *V/L representation for language and vision* can be independently derived by the learned modality encoders whereas their fused representation can represent the *V-L representation for joint modality*.

4 Pretraining

In this section, we review the pretraining regarding the datasets, tasks and objective designs.

4.1 Datasets

Conceptual Captions (CC, roughly 3M) and SBU Captions (SBU, around 1M) of enormous size and diversified nature are the most commonly used webly collected datasets for VLPM pretraining [Lu *et al.*, 2019; Zhou *et al.*, 2020; Li *et al.*, 2021b; Li *et al.*, 2020a; Fei *et al.*, 2021]. It is found that a larger sized corpus leads to better performance in downstream transfer tasks [Lu *et al.*, 2019; Wen *et al.*, 2021; Jia *et al.*, 2021]. Some simple Dual-encoder VLPMs [Radford *et al.*, 2021; Jia *et al.*, 2021] collect even larger scaled corpus from the web, such as WIT (400M) [Radford *et al.*, 2021] and ALIGN(1.8B) [Jia *et al.*, 2021].

Another combination that leads to better domain adaptation is the *out-of-domain+in-domain* datasets, where they define the web-based CC/SBU as *out-of-domain* and the MS-COCO (COCO)/Visual Genome (VG) as the *in-domain* datasets because most downstream tasks are built on them [Chen *et al.*, 2020; Yu *et al.*, 2021; Sun *et al.*, 2021; Kim *et al.*, 2021; Wang *et al.*, 2021a; Li *et al.*, 2021a]. The in-domain datasets can also be the *task-specific* datasets that specifically come from the commonly evaluated downstream tasks (e.g., GQA/VQA2.0) [Lu *et al.*, 2020; Xia *et al.*, 2021; Zhang *et al.*, 2021; Li *et al.*, 2020b; Wen *et al.*, 2021].

Some VLPMs target specialized tasks/domains such as Visual Dialogue (VD) and pretrain on only the *domain/task-specific* dataset due to their reduced suitability to the aforementioned text-image datasets regarding domain nature or dataset format [Gao *et al.*, 2020; Zhuge *et al.*, 2021; Wang *et al.*, 2020; Hao *et al.*, 2020]. There are also efforts of jointly pretraining with single-modal tasks on auxiliary *single-modality* data source, i.e., non-paired image or text collection, for reinforcing single-modal representations [Su *et al.*, 2019; Li *et al.*, 2021a; Fei *et al.*, 2021; Ni *et al.*, 2021; Singh *et al.*, 2021].

4.2 Tasks and Objectives

There are three most commonly used *cross-modal pretraining tasks*[1]: **1) Cross-modal Masked Language Modeling (CMLM)** extends the Masked Language Modeling (MLM) from Bert pretraining [Devlin *et al.*, 2019] to multi-modal setting for learning the contextualized V-L representation, utilizing the bi-directional *attention-based V-LIM*. This is proved to be helpful for transferring the pretrained language model Bert into multi-modal setting [Wang *et al.*, 2020] and thus makes it one of the most essential VLPM pretraining tasks [Su *et al.*, 2019; Huang *et al.*, 2020; Chen *et al.*, 2020]. The task goal is to predict the masked tokens in text sequence based on the observation of their surrounding context, including both the unmasked textual tokens and all the visual tokens, by minimising the negative log-likelihood (NLL) loss. While most VLPMs mask out WordPiece sub-tokens as in Bert, some achieve better transfer performance via masking over semantically integral token groups such as complete word [Kim *et al.*, 2021; Gao *et al.*, 2020] or segment [Yu *et al.*, 2021; Li *et al.*, 2021a; Li *et al.*, 2020b].

2) Cross-modal Masked Region Modeling (CMRM) is a vision-supervised counterpart of CMLM, initially proposed by RoI-based VLPMs. It randomly masks out tokens in the visual sequence instead [Chen *et al.*, 2020; Tan and Bansal, 2019; Lu *et al.*, 2019; Li *et al.*, 2020a; Su *et al.*, 2019]. Thus far, there exist three variant objectives: **a) Region Label Classification (CMRM$_\mathbf{C}$)** predicts the object class of each masked region via minimizing the cross-entropy (CE) loss calculated based on the one-hot encoded object class from the object detector as ground-truth and the normalized distribution of the VLPM prediction over the possible object classes. To mitigate the potential classification error of the object detector, **b) Label Distribution Approximation (CMRM$_\mathbf{D}$)** uses the probability distribution of the object class as a soft supervision by minimising the KL divergence loss between the object class distribution from the object detector and the normalized VLPM prediction distribution. Comparatively, **c) Region Feature Regression (CMRM$_\mathbf{R}$)** learns to regress the VLPM output of each masked region into its input feature derived from the object detector using the L2 loss. It is commonly applied together with CMRM$_\mathrm{C}$ [Tan and Bansal, 2019; Ni *et al.*, 2021; Li *et al.*, 2021a] or CMRM$_\mathrm{D}$ [Sun *et al.*, 2021; Chen *et al.*, 2020] for more robust visual content modeling and joint cross-modal learning. Especially, when the visual token is represented at *patch* level instead of RoI, CMRM can also be extended to masked *patch* modeling [Huang *et al.*, 2021; Zhuge *et al.*, 2021; Gao *et al.*, 2020].

VLPMs with *attention-based V-LIM* commonly treat **3) Cross-modal Alignment (CA)** as a binary classification problem, aiming to predict whether the input image-text pair is semantically matched or not based on the global V-L representation, by applying binary cross-entropy loss. The negative pairs can be sampled by randomly replacing either the image/text in the positive pair with another image/text from other data pairs in the corpus. On the other hand, dual-encoder based VLPMs with *VSE-based V-LIM* all regard CA as a ranking problem for finding the best match and apply the contrastive learning optimization objective [Radford *et al.*, 2021; Sun *et al.*, 2021; Jia *et al.*, 2021; Wen *et al.*, 2021]. They learn the decoupled V/L representation in the semantically shared VSE space. Most of them adopt in-batch negatives for convenience, which simply treats the text/image from the remaining pairs within the batch as negatives. In addition to the two *global-level alignment* CA objectives above, there are also several *fine-grained alignment* variants [Chen *et al.*, 2020; Kim *et al.*, 2021] that try to enforce the matching between the fine-grained components such as visual regions/patches and textual words of the image-

[1] Task/objective names may vary slightly in different papers.

text pairs. One essential difference between the global-level and the fine-grained objectives is that the former introduces negative samples, which is found harmful for the fusion-based VLPMs with attention-based V-LIMs [Su *et al.*, 2019]. Thus, some VLPMs [Li *et al.*, 2021b; Singh *et al.*, 2021; Wang *et al.*, 2021a] introduce the single-modal encoders for its global-level CA while simultaneously keeping the fusion-based encoders for other tasks, to combine the advantages of the dual-encoder-based and fusion-based structures.

However, these three tasks cannot fulfil the goal of V-L generation tasks, such as Image Generation (IC) [Zhou *et al.*, 2020; Xia *et al.*, 2021; Li *et al.*, 2021b; Wang *et al.*, 2021b]. Thus, encoder-decoder-based VLPMs especially design the pretraining tasks to involve a decoding process. Those unified encoder-decoder VLPMs simply extend the bidirectional CMLM to **seq2seq CMLM** as an auxiliary pretraining task for text generation [Zhou *et al.*, 2020; Li *et al.*, 2021a; Wang *et al.*, 2020]. Other VLPMs with decoupled encoder-decoder emphasize the inherently different peculiarities of the understanding and generation disciplines, where the former entails the unrestricted information passing across modalities whereas the latter only involves visual-to-textual information passing. They keep the bi-directional cross-modal modeling only to the encoder while generating the text via a separate decoder in an auto-regressive manner [Xia *et al.*, 2021; Li *et al.*, 2021b; Wang *et al.*, 2021b].

From another perspective, applying downstream task such as the **IC** above for pretraining can be regarded as a kind of *downstream-driven pretraining task* [Tan and Bansal, 2019; Gao *et al.*, 2020]. Some special pretraining tasks especially designed for the specific downstream task/domain can also be downstream-driven [Wang *et al.*, 2020; Murahari *et al.*, 2020; Hao *et al.*, 2020; Yang *et al.*, 2021], e.g., the Answer Prediction (**AnP**) for the Visual Dialogue(VD) task [Wang *et al.*, 2020]. Besides, there are VLPMs that jointly pretrain on additional text and/or image collection with the single-modal tasks [Su *et al.*, 2019; Wang *et al.*, 2021a; Wen *et al.*, 2021; Li *et al.*, 2021a], including the single-modal multi-lingual settings [Ni *et al.*, 2021; Fei *et al.*, 2021].

5 Transfer Learning and Evaluation

In this section, we will introduce the downstream Visual-Linguistic understanding (V-L Understanding), Visual-Linguistic generation (V-L Generation) and single modal tasks that VLPMs mainly focus on.

V-L Understanding tasks refer to the tasks that require a model to capture both visual and linguistic semantics from the input and learn the correspondence between them. Most VLPMs apply a variety of V-L Understanding tasks for evaluation, while some just focus on a specific domain. Table 1 summarises the downstream tasks in V-L Understanding.

Visual Question Answering (VQA). Most works [Tan and Bansal, 2019; Lu *et al.*, 2019; Zhou *et al.*, 2020; Su *et al.*, 2019; Gan *et al.*, 2020; Wang *et al.*, 2021a] formalize it as a classification problem to select an answer from a pool of common answers. However, SimVLM [Wang *et al.*, 2021b] tries to decode open-ended answers in an auto-regressive manner to remove the limit posed by the fixed answer vocabulary. Among these models, Self-attention-based VLPMs work better than Co-attention-based VLPMs, and SimVLM outperforms all other VLPMs on the VQA v2.0 dataset.

Cross Modal Retrieval (CMR). Some VLPMs [Li *et al.*, 2020b; Gao *et al.*, 2020] treat it as a binary classification task to predict whether each image-caption pair is matching, while others [Chen *et al.*, 2020; Li *et al.*, 2021b; Wen *et al.*, 2021; Li *et al.*, 2020a] solve it as a ranking problem to learn a similarity score to be maximized between positive pairs and minimized between negative pairs with a contrastive or CE loss.

Text Classification. Some VLPMs [Tan and Bansal, 2019; Gan *et al.*, 2020; Lu *et al.*, 2020; Huang *et al.*, 2020; Chen *et al.*, 2020] work on the *Natural Language for Visual Reasoning (NLVR)* task to perform binary classification based on special token representations, and some works [Gan *et al.*, 2020; Lu *et al.*, 2020; Chen *et al.*, 2020; Li *et al.*, 2021a; Huang *et al.*, 2021; Kim *et al.*, 2021] focus on the *Visual Entailment (VE)* problem to evaluate their methods.

Visual Commonsense Reasoning (VCR). The VCR task is usually divided into two sub-tasks to predict an answer and a rationale separately, and each sub-task is similar to answering a multiple-choice question out of four options. Therefore, most VLPMs [Li *et al.*, 2019; Lu *et al.*, 2019; Su *et al.*, 2019] formulate each sub-task as a binary classification problem for each option, or a 4-way classification problem as done by Unicoder-VL [Li *et al.*, 2020a].

Referring Expression Comprehension (REC). VLPMs for REC usually perform a binary classification on each image region for whether it is the target of the phrase and select the region with the highest score at inference stage [Lu *et al.*, 2019; Su *et al.*, 2019; Gan *et al.*, 2020; Lu *et al.*, 2020; Chen *et al.*, 2020; Yu *et al.*, 2021]. The VLPMs pretrained with in-domain datasets[Chen *et al.*, 2020; Gan *et al.*, 2020] generally work better on this task than those pretrained with only out-of-domain datasets.

Visual Relationship Detection (VRD). Only RVL-BERT focuses on this domain with two formulations [Chiou *et al.*, 2021]. It can either be formulated as a ranking problem with the goal of ranking all possible subject-predicate-object triplets between any pair of object regions or as a binary classification problem to predict if a given subject-predicate-object triplet holds.

Visual Dialogue (VD). This problem can be formulated in a discriminative manner to select an answer from 100 candidates as done by VisDialBERT [Murahari *et al.*, 2020] and VDBERT [Wang *et al.*, 2020], or in a generative manner to auto-regressively decode the answer as done by VD-BERT [Wang *et al.*, 2020].

Visual Linguistic Navigation (VLN). Some VLPMs [Majumdar *et al.*, 2020] try to solve this task in a discriminative setting by selecting the correct path among one positive path and several negative sample paths. In contrast, other VLPMs [Hong *et al.*, 2021] try to solve it in a generative setting to make an action in between each state of the path by applying a reinforcement learning objective together with an imitation learning objective.

Tasks	Papers
VQA	[Li *et al.*, 2019], [Tan and Bansal, 2019], [Lu *et al.*, 2019], [Zhou *et al.*, 2020], [Su *et al.*, 2019], [Gan *et al.*, 2020], [Lu *et al.*, 2020], [Huang *et al.*, 2020], [Chen *et al.*, 2020], [Li *et al.*, 2020b], [Yu *et al.*, 2021], [Li *et al.*, 2021b], [Zhang *et al.*, 2021], [Huang *et al.*, 2021], [Zhuge *et al.*, 2021], [Wang *et al.*, 2021b], [Li *et al.*, 2021a], [Wang *et al.*, 2021a], [Kim *et al.*, 2021], [Singh *et al.*, 2021]
CMR	[Lu *et al.*, 2019], [Li *et al.*, 2020a], [Gan *et al.*, 2020], [Lu *et al.*, 2020], [Huang *et al.*, 2020], [Gao *et al.*, 2020], [Chen *et al.*, 2020], [Li *et al.*, 2020b], [Yu *et al.*, 2021], [Li *et al.*, 2021b], [Zhang *et al.*, 2021], [Ni *et al.*, 2021], [Sun *et al.*, 2021], [Huang *et al.*, 2021], [Li *et al.*, 2021a], [Wen *et al.*, 2021], [Fei *et al.*, 2021], [Jia *et al.*, 2021], [Wang *et al.*, 2021a], [Kim *et al.*, 2021], [Singh *et al.*, 2021]
NLVR	[Li *et al.*, 2019], [Tan and Bansal, 2019], [Gan *et al.*, 2020], [Lu *et al.*, 2020], [Huang *et al.*, 2020], [Chen *et al.*, 2020], [Li *et al.*, 2020b], [Zhang *et al.*, 2021], [Huang *et al.*, 2021], [Wang *et al.*, 2021b], [Wang *et al.*, 2021a], [Kim *et al.*, 2021]
VE	[Gan *et al.*, 2020], [Lu *et al.*, 2020], [Chen *et al.*, 2020], [Li *et al.*, 2021a], [Huang *et al.*, 2021], [Wang *et al.*, 2021b], [Kim *et al.*, 2021], [Singh *et al.*, 2021]
VCR	[Li *et al.*, 2019], [Lu *et al.*, 2019], [Li *et al.*, 2020a], [Su *et al.*, 2019], [Gan *et al.*, 2020], [Chen *et al.*, 2020], [Yu *et al.*, 2021], [Li *et al.*, 2021b]
REC	[Lu *et al.*, 2019], [Su *et al.*, 2019], [Gan *et al.*, 2020], [Lu *et al.*, 2020], [Chen *et al.*, 2020], [Yu *et al.*, 2021]
VRD	[Chiou *et al.*, 2021]
VD	[Wang *et al.*, 2020], [Murahari *et al.*, 2020]
VLN	[Majumdar *et al.*, 2020], [Hao *et al.*, 2020], [Hong *et al.*, 2021]

Table 1: Categories of Downstream Tasks in Visual Language Understanding

V-L Generation tasks generate texts/images when the other modality is included in the input. Most generative VLPMs [Zhou *et al.*, 2020; Xia *et al.*, 2021; Li *et al.*, 2020b; Zhang *et al.*, 2021; Li *et al.*, 2021a; Li *et al.*, 2021b; Wang *et al.*, 2021b; Hu *et al.*, 2021] focus on **Image Captioning (IC)** task on COCO Captioning dataset by minimizing the NLL loss. Some of them [Li *et al.*, 2020b; Zhang *et al.*, 2021; Wang *et al.*, 2021b; Hu *et al.*, 2021] try the Nocaps task, which generates captions for images containing open domain novel objects. In addition, KaleidoBERT [Zhuge *et al.*, 2021] extends this task to the fashion domain, and TAP [Yang *et al.*, 2021] focuses on the TextCaps task (and the Text VQA task), which requires models to understand scene texts by applying CE loss at multiple decoding steps. VLPMs pretrained with extra in-domain datasets generally perform better except on the Nocaps. Similarly, multi-lingual VLPMs [Wang *et al.*, 2021b] focus on the **Multi-modal Machine Translation (MMT)** task by minimizing the NLL loss.

Despite focusing on multi-modal pretraining, some VLPMs also investigate how their models can generalize to single modal tasks. SimVLM [Wang *et al.*, 2021b] and ALIGN [Jia *et al.*, 2021] perform the image classification task of predicting object categories, and KaleidoBERT [Zhuge *et al.*, 2021] does it similarly to predict the category and subcategory of commercial products. Some VLPMs [Wang *et al.*, 2021b; Li *et al.*, 2021a] try to evaluate the natural language understanding capability of their models based on the GLUE benchmark tasks (or its subset). An improvement in linguistic tasks has been observed when pretrained with in-domain datasets, even with smaller pretraining data sizes.

6 Conclusion and Future Research Directions

This paper presents an overview of the recent advances in VLPMs for producing joint representations of visions and languages, images and text pairs. We mainly summarise vision and language input encoding methods and mainstream VLPMs. We also discuss several useful pretraining and fine-tuning tasks and strategies. We do hope that the paper will provide insightful guidance for both CV and NLP researchers who work on joint and cross-modal learning. To advance this, there are several promising future directions for VLPMs.

V-L Interaction Modeling. Although various VL interaction model extensions have been proposed in Section 3, there is still a significant challenge in aligning vision and language content. Most pretrained models focus on masking at the task level or input level, which does not directly align the features between image and text. Incorporating masking strategy at the embedding level has been shown to be effective [Zhuge *et al.*, 2021]. It is promising to investigate how to explicitly align the embedding features between image and text so that it can learn fine-grain representations.

VLPM Pretraining Strategy. There still lacks systematic experiments and analysis on V-L-based multi-tasking synergy for VLPM pretraining. Few VLPMs explore multi-stage training [Li *et al.*, 2019; Wang *et al.*, 2020; Hong *et al.*, 2021; Xia *et al.*, 2021; Chen *et al.*, 2020], and 12-in-1 [Lu *et al.*, 2020] is the only one that tries to group the tasks of similar natures and test the performance boost from intra- and inter-group pretraining perspectives. It faces significant challenges to finding the answer in one step because it entails multiple factors such as selecting datasets (both multi-modal and single-modal), task design, task grouping, and order (multi-stage). Moreover, the effectiveness of the pretraining process may vary for different downstream targeted tasks. However, it is still worth exploring, step by step, how the V-L based multi-tasking can be implemented for VLPM pretraining that can generate the best transfer performance on the specific targeted domain/tasks, which will provide promising and valuable guidelines for the future development of VLPMs.

Training Evaluation. The trained VLPMs can be only evaluated during the downstream tasks. This may waste a lot of computation cost when the models are defective. It is worth exploring some metrics, such as perplexity, during the training procedure. Hence, we can guarantee the performance of the trained VLPMs in advance.

References

[Bugliarello *et al.*, 2021] Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language berts. *Transactions of the Association for Computational Linguistics*, 9:978–994, 2021.

[Chen *et al.*, 2020] 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.

[Chiou *et al.*, 2021] Meng-Jiun Chiou, Roger Zimmermann, and Jiashi Feng. Visual relationship detection with visual-linguistic knowledge from multimodal representations. *IEEE Access*, 9:50441–50451, 2021.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, 2019.

[Dosovitskiy *et al.*, 2020] 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.

[Fei *et al.*, 2021] Hongliang Fei, Tan Yu, and Ping Li. Cross-lingual cross-modal pretraining for multimodal retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

[Gan *et al.*, 2020] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020.

[Gao *et al.*, 2020] 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.

[Hao *et al.*, 2020] 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.

[Hong *et al.*, 2021] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1643–1653, 2021.

[Hu *et al.*, 2021] Xiaowei Hu, Xi Yin, Kevin Lin, Lei Zhang, Jianfeng Gao, Lijuan Wang, and Zicheng Liu. Vivo: Visual vocabulary pre-training for novel object captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1575–1583, May 2021.

[Huang *et al.*, 2020] 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.

[Huang *et al.*, 2021] 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.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

[Kim *et al.*, 2021] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*, 2021.

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

[Li *et al.*, 2020a] 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.

[Li *et al.*, 2020b] 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.

[Li *et al.*, 2021a] 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. 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 2592–2607, 2021.

[Li *et al.*, 2021b] Yehao Li, Yingwei Pan, Ting Yao, Jingwen Chen, and Tao Mei. Scheduled sampling in vision-language pretraining with decoupled encoder-decoder network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8518–8526, 2021.

[Lu *et al.*, 2019] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in NeurIPS*, 32:13–23, 2019.

[Lu *et al.*, 2020] 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.

[Majumdar *et al.*, 2020] Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. Improving vision-and-language navigation with image-text pairs from the web. In *European Conference on Computer Vision*, pages 259–274. Springer, 2020.

[Mogadala *et al.*, 2021] Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *Journal of Artificial Intelligence Research*, 71:1183–1317, 2021.

[Murahari *et al.*, 2020] 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.

[Ni *et al.*, 2021] 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.

[Radford *et al.*, 2021] 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. *arXiv preprint arXiv:2103.00020*, 2021.

[Singh *et al.*, 2021] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. *arXiv preprint arXiv:2112.04482*, 2021.

[Su *et al.*, 2019] 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.

[Sun *et al.*, 2021] Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 982–997, 2021.

[Tan and Bansal, 2019] 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.

[Wang *et al.*, 2020] Yue Wang, Shafiq Joty, Michael Lyu, Irwin King, Caiming Xiong, and Steven CH Hoi. Vd-bert: A unified vision and dialog transformer with bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3325–3338, 2020.

[Wang *et al.*, 2021a] 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.

[Wang *et al.*, 2021b] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[Wen *et al.*, 2021] Keyu Wen, Jin Xia, Yuanyuan Huang, Linyang Li, Jiayan Xu, and Jie Shao. Cookie: Contrastive cross-modal knowledge sharing pre-training for vision-language representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2208–2217, 2021.

[Xia *et al.*, 2021] 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.

[Yang *et al.*, 2021] Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. Tap: Text-aware pre-training for text-vqa and text-caption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8751–8761, 2021.

[Yu *et al.*, 2021] 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.

[Zhang *et al.*, 2021] 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.

[Zhou *et al.*, 2020] 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.

[Zhuge *et al.*, 2021] 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.