Kaleido-BERT: Vision-Language Pre-training on Fashion Domain

Mingchen Zhuge[1,†] Dehong Gao[1,†] Deng-Ping Fan[2,✉]
Linbo Jin[1] Ben Chen[1] Haoming Zhou[1] Minghui Qiu[1] Ling Shao[2]
[1] Alibaba Group [2] Inception Institute of AI (IIAI)
http://dpfan.net/Kaleido-BERT

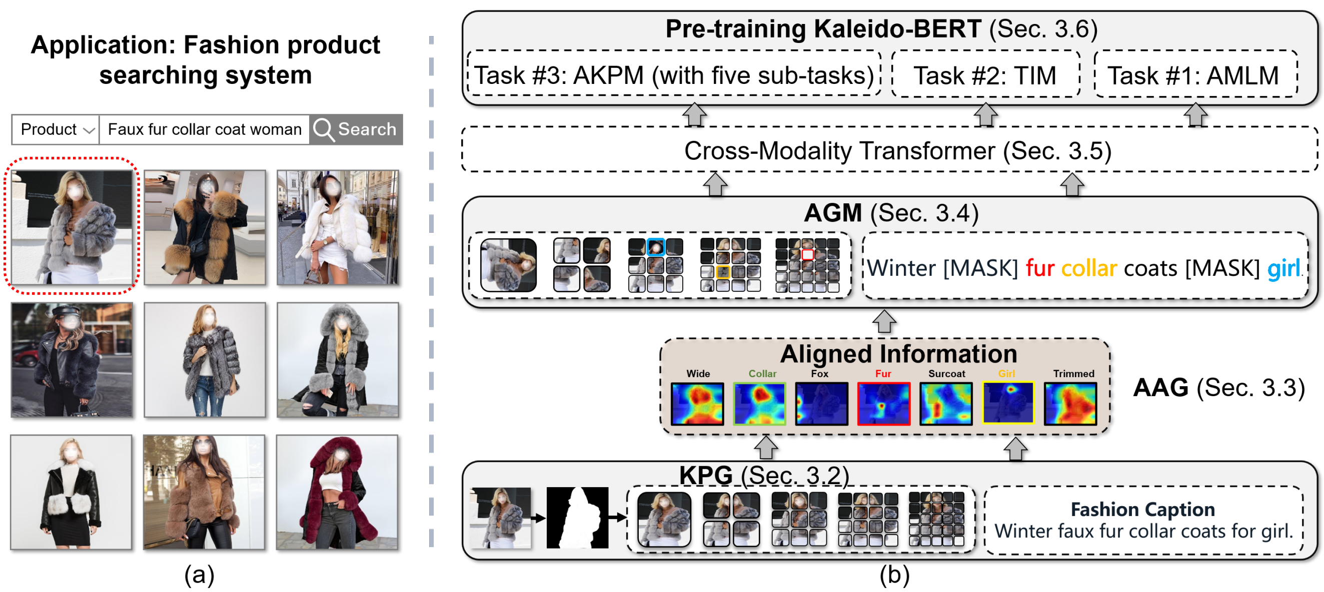

Figure 1: **Vision-Language (VL) pre-training architecture on fashion.** We propose a novel VL pre-training architecture (**Kaleido-BERT**), which consists of a Kaleido Patch Generator (KPG), Attention-based Alignment Generator (AAG), and Alignment Guided Masking (AGM) strategy to learn better VL feature embeddings. Kaleido-BERT achieves the state-of-the-art on the standard public Fashion-Gen dataset and deploys to the online system (a).

Abstract

*We present a new vision-language (VL) pre-training model dubbed **Kaleido-BERT** , which introduces a novel kaleido strategy for fashion cross-modality representations from transformers. In contrast to random masking strategy of recent VL models, we design alignment guided masking to jointly focus more on image-text semantic relations. To this end, we carry out five novel tasks,* i.e.*, rotation, jigsaw, camouflage, grey-to-color, and blank-to-color for self-supervised VL pre-training at patches of different scale. Kaleido-BERT is conceptually simple and easy to extend to the existing BERT framework, it attains state-of-the-art results by large margins on four downstream tasks, including text retrieval (R@1: 4.03% absolute improvement), image retrieval (R@1: 7.13% abs imv.), category recognition (ACC: 3.28% abs imv.), and fashion captioning (Bleu4: 1.2 abs imv.). We validate the efficiency of Kaleido-BERT on a wide range of e-commerical websites, demonstrating its broader potential in real-world applications.*

1. Introduction

Transformers [14, 68], first designed for Natural Language Processing (NLP), have achieved great success in a number of other areas as well [5, 11], including the vision (*e.g.*, Selfie [66], DETR [6], ViT [34], and PVT [69]) and vision-language (ViLBERT [45], VL-BERT [60], OSCAR [42]) communities. However, for VL Pre-Training Model (PTM), current approaches, such as VL-BERT [60] and UNITER [9] focus on learning text and image representation of a general domain (*i.e.*, coarse matching). As such, these techniques will benefit for general cross-modality representation learning.

† Equal; * Corresponding author: Deng-Ping Fan (dengpfan@gmail.com).

No.	Pre-Training Model	Year	Pub.	Architecture	Training Set	Core Idea	Domain	Pre-train	Finetune	Vision Feature	Code
1	VisualBERT [40]	2019	arXiv	1-Stream	Coco	First Image-Text PTM *AFAK*	Image	✓	U	RoI	Torch
2	CBT [63]	2019	arXiv	2-Stream	Kinetics [30]	Noise contrastive estimation loss	Video	✓	U/G/O	Frame	N/A
3	VideoBERT [64]	2019	ICCV	1-Stream	SC	First Vedio-Text PTM *AFAK*	Video	✓	G/O	Frame	N/A
4	B2T2 [2]	2019	EMNLP	1-Stream	CC	Explicitly reference RoI to text before inputting PTM	Image	✓	U	RoI	Tensorflow
5	LXMERT [38]	2019	EMNLP	2-Stream	VG+Coco	Three encoders for ROIs, language, and cross-modality features	Image	✓	U	RoI	Torch
6	ViLBERT [45]	2019	NeurlIPS	2-Stream	CC	Cross-modality co-attention layers	Image	✓	U	RoI	Torch
7	ImageBERT [55]	2020	arXiv	1-Stream	CC+VG+SC	Pre-training with a large-scale Image-Text dataset	Image	✓	U	RoI	N/A
8	Unicoder-VL [38]	2020	AAAI	1-Stream	CC+SBU	Masked object classification	Image	✓	U	RoI	N/A
9	VLP [82]	2020	AAAI	1-Stream	CC	Unified en/decoder PTM for VL generation and understanding	Image	✓	U/G	RoI	Torch
10	VL-BERT [61]	2020	ICLR	1-Stream	CC	Visual feature embedding	Image	✓	U	RoI	Torch
11	VD-BERT [70]	2020	EMNLP	1-Stream	VisDial [12]	Video-Dialog pre-training	Video	✓	O	RoI	Torch
12	VLN-BERT [48]	2020	ECCV	2-Stream	Matterport3D [7]	Path selection in VL navigation	Image		O	RoI	N/A
13	HERO [39]	2020	EMNLP	3-Stream	TV+HT100M	Video-subtitle matching & Frame order modeling	Video	✓	U	Frame	N/A
14	XGPT [73]	2020	arXiv	1-Stream	CC+SC	Improve VL generative ability	Image	✓	U/G	RoI	N/A
15	InterBERT [43]	2020	arXiv	1-Stream	Coco+CC+SBU	Masked group modeling	Image	✓	U	RoI	Torch
16	VILLA [20]	2020	NeurlIPS	2-Stream	Coco+CC+SBU	Adversarial pre-training and finetune	Image	✓	U/O	RoI	Torch
17	ActBERT [83]	2020	CVPR	1-Stream	HT100M	Global frame and local object regions & Tangled transformer	Video	✓	U/O	Frame & RoI	N/A
18	PREVALENT [24]	2020	CVPR	2-Stream	Matterport3D [7]	Pre-train with image-text-action triplets	Image	✓	O	Image	Caffe & C++
19	12-IN-1 [46]	2020	CVPR	2-Stream	ES	Multi-task training	Image	✓	U	RoI	Torch
20	Pixel-BERT [27]	2020	arXiv	1-Stream	Coco+VG	Pixel-level VL semantics alignment	Image	✓	U	Pixel	N/A
21	FashionBERT [21]	2020	SIGIR	1-Stream	Fashion-Gen [58]	Patches & Adaptive loss	Image	✓	U	Patch	Tensorflow
22	UNITER [9]	2020	ECCV	1-Stream	Coco+VG+CC+SBU	Conditional masking & Word region alignment	Image	✓	U	RoI	Torch
23	VisDial-BERT [50]	2020	ECCV	2-Stream	CC+VQA [4]	Adapt ViLBERT for visual dialog	Image		O	RoI	Torch
24	OSCAR [42]	2020	ECCV	1-Stream	ES	Object tags as anchor points	Image	✓	U/G	RoI	Torch
25	ERNIEL-VIL [78]	2020	arXiv	2-Stream	CC+SBU	VL PTM with knowledge-enhanced ERNIE [80]	Image	✓	U	RoI	Paddle
26	RVL-BERT [10]	2020	arXiv	1-Stream	VDR [44]	Visual relationship detection with VL-BERT	Image		U	RoI	Torch
27	UniVL [47]	2020	arXiv	2-Stream	HT100M	Five pre-training objectives & Two pre-training strategies	Vedio	✓	U/G	Frame	N/A
28	MMFT-BERT [32]	2020	EMNLP	3-Stream	TV	Multi-modal fusion PTM	Image	✓	U	RoI	Torch
29	**Kaleido-BERT (OUR)**	2021	CVPR	1-Stream	Fashion-Gen [58]	Kaleido patches & Pre-alignment with masking	Image	✓	U/G	Patch & Coordinate	Tensorflow

Table 1: **Summary of 28 previous representative cross-modality methods and our Kaleido-BERT model. Training Set:** Coco = *MSCOCO Caption* [8]. VG = *Visual Genome* [35]. CC = *Conceptual Caption* [59]. SBU = *SBU Captions* [53]. TV = *TVQA* [37]. HT100M = *HowTo100M* [49]. SC: *Self Collection*. ES: *12-in-1 and OSCAR ensemble 12, 5+ datasets, respectively.* **Finetune:** U = Understanding tasks (*e.g*. classification). G = Generation tasks (*e.g*. image caption). O = Others (*e.g*. action task).

However, in the various e-commercial situations (*e.g*., accessories, clothing, toys), the main goal is to learn the *fine-grained* representation (*e.g*. short sleeve, cotton and jersey) rather than only the *coarse* representation (what, where) in the general domain. In this case, the current general VL models [9, 60] are sub-optimal for fashion-based tasks [1, 26, 67], and could be unfavorable when deploying global features based models to attribute-aware tasks, such as searching for a specific fashion captioning [75] and fashion catalog/object [15], where it is essential to extract fine-grained features or similarities [65] from image and text.

In this paper, we propose a novel framework (see Fig. 1) for the fashion-based tasks. The core idea is to focus on fine-grained representation learning and to bridge the semantic gaps between text and image. To achieve this goal, we first introduce an efficient "*kaleido*" strategy, which extracts a series of multi-grained image patches for the image-modality. As a result, our model is named as **Kaleido-BERT**. This strategy is scalable and largely alleviates the aforementioned coarse presentation issue by introducing the patch-variant pre-training scheme. Furthermore, to bridge the semantic gap between different modalities, attention mechanism is employed to build pre-alignments between kaleido patches and text tokens. This pre-alignment information further guides the masking strategy for pre-training. Kaleido-BERT[1] is forced to *explicitly* learn semantic information across modalities. In summary, our contributions are as follows:

- **Kaleido Patch Generator**: We propose the kaleido strategy to generate a kaleido of multi-grained features. With the related pre-training tasks, *i.e*. rotation, jigsaw, camouflage, grey-to-color, and blank-to-color, Kaleido-BERT learns fine-grained cross-modality information and outperforms the fixed-patch or RoI-based VL models in the fashion domain.
- **Attention-based Alignment Generator**: Kaleido-BERT introduces the pre-alignment strategy to infer a cross-modality mapping between kaleido patches and text tokens. These pre-alignment pairs largely fill the semantic gaps between modalities.
- **Alignment Guided Masking**: We present an alignment-guided masking strategy to explicitly force Kaleido-BERT to learn the semantic connections between vision and language. Experiments show the importance of the attention-based pre-alignment and the alignment masking strategy.

2. Related Work

There is a large body of VL modeling literature [3, 4, 23, 28, 31, 54, 62, 77, 79], and we briefly introduce the transformer-based methods in this section. More detailed summary can be found in Tab. 1.

2.1. Vision-Language Pre-training

Recent transformer-based pre-training frameworks, such as BERT [14], GPT2 [57], XLNet [76], and GPT3 [5], have revolutionized NLP tasks. Motivated by these studies, many cross-modal pre-training models for vision-language (*e.g*., video/image and text pairs) have been designed. For video-text pair models, CBT [63] and VideoBERT [64] are pioneering work that study the capability of pre-training learning. ActBERT [83] and HERO [39] focus more on down-

[1] https://github.com/mczhuge/Kaleido-BERT/.

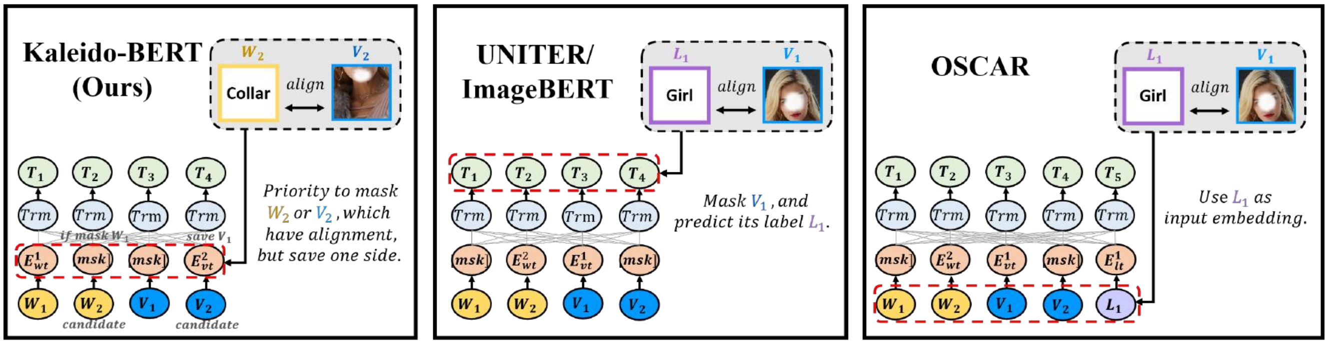

Figure 2: **Different utilization of alignment information in VL pre-training architectures. T** = *Task*. **W** = *Word*. **V** = *Visual Token*. **L** = *Object Detection Label*. **E** = *Embedding*. **Trm** = *Transformer Block*.

stream applications, while UniVL [47] focuses on both video-language understanding and generation tasks.

For image-text pair models, they can be categorized into single-stream [2, 9, 10, 21, 27, 38, 40–43, 55, 61, 70, 73, 82] and two-stream [13, 15, 24, 38, 45, 48, 50, 78] or even three-stream [32] according to the network architecture of the single-modal input. In single-stream models, the features of different modalities are directly fed into a Transformer. In contrast, in two-stream models, they are first processed by two single-modal networks before fed into a Transformer, and so forth in three-stream models. ViLBERT [45] claims that the two-stream structure is superior to the single-stream, while VL-BERT [61] finds that the single-stream models achieve more promising results, as these models have more cross-modality information interactions. VisualBERT [40] and B2T2 [2] are single-stream models and derive a unified VL understanding network. With the same spirit of focusing on generic VL tasks, many concurrent models, *e.g*., Unicoder-VL [38], VLP [82], ViLBERT [45], VL-BERT [61], have been proposed under the BERT framework. In contrast to the boom in generic tasks (*e.g*., VCR [9, 43, 78], VQA [32, 38]), other tasks such as visual relationship detection (RVL-BERT [10]), visual navigation (*i.e*., PERVALENT [24] and VLN-BERT [48]), and visual dialog (*e.g*., VisualD [50], VD-BERT [70]) are still in their infancy. More recently, Lu *et al*. [46] shows that multi-task VL learning can lead to a significant improvement over isolated task learning. Similarly, OSCAR [42] achieves new state-of-the-art performance on many representative VL tasks (*e.g*., image captioning like XGPT [73], image-text retrieval like Image-BERT [55]).

Advances have also been made by the VirTex [13] model in image classification, object detection, and instance segmentation filed, by using semantically dense captions to learn visual representations. Another notable study is the recent ACL work [41] in which the authors creatively demonstrate that the attention head of the VL model can perform entity grounding and syntactic grounding. Unlike all the above-mentioned works, Pixel-BERT [27] considers aligning vision-language features at a pixel level instead of using region-based image features.

As shown in Fig. 2, our Kaleido-BERT focuses on a masking strategy at the embedding level rather than at the task level (*e.g*., LXMERT [38] and UNITER [9]) or input level such as OSCAR [42]. Kaleido-BERT *explicitly* align the embedding features between image and text so that it can learn fine-grain representations for fashion tasks.

2.2. Fashion-Based Task

As described in §. 2.1, existing VL models mainly focus on relatively *coarse representations*, while less attention has been paid to fine-grained representation learning for the fashion-based task. There are two concurrent studies [15, 21] resembling our work. FashionBERT [21] was the first published work in the fashion domain. The concurrent work, MAAF [15], aims to derive a modality-agnostic attention fusion strategy to address the undifferentiated text and image query task. Unlike FashionBERT, which utilizes a patch-fixed masking strategy, the MAAF adopts an image-level attention mechanism. We argue that these two schemes restrict the power of the pre-trained representation learning, especially for the fine-grained fashion task. As a consequence, a more flexible solution with patch-variant is urgently required.

To the best of our knowledge, the proposed Kaleido-BERT is the first to present the effectiveness of alignment guided masking by jointly focusing more on image-text coherence for the fashion domain.

3. Proposed Kaleido-BERT

In this section, we introduce our Kaleido-BERT, which learns the *fine-grained* VL features for the fashion domain rather than the *coarse representation* features for VL tasks. We adopt the standard transformer designed for NLP to make our Kaleido-BERT scalable over a varying number of transformer-based VL learning task.

3.1. Model Overview

The architecture of our Kaleido-BERT is illustrated in Fig. 1. There are five stages: (1) Kaleido-BERT takes two inputs: a text (*e.g.*, image caption or description) and corresponding image patches generated by our Kaleido Patches Generator (**KPG**). Similar to LXMERT [38], each text is represented as a sequence of tokens and each image is represented as a sequence of kaleido patches. (2) At the embedding stage, we propose the Attention-based Alignment Generator (**AAG**) to generate pre-alignments between text tokens and kaleido patches so that the image and text are explicitly aligned semantically. (3) Different from existing random masking strategy, we proposed to adopt an Alignment Guided Masking (**AGM**) strategy to relieve the difficulty of cross-modality modeling. (4) Text tokens and kaleido patches fully interact in Kaleido-BERT, which gradually learns VL semantic information and produces the cross-modality fine-grained representations. (5) Finally, we adopt five new kaleido tasks (*i.e.*, rotation, jigsaw, camouflaged, grey-to-color and blank-to-color tasks) besides the masked language modeling and image-text matching tasks to supervise the network. Our implementation is based on the `EasyTransfer` [2]/`Huggingface` [3] library. We refer the readers to this de facto standard library for details.

3.2. Kaleido Patch Generator

Given an image as input, we obtain the multi-grained patches by the proposed Kaleido Patch Generator (**KPG**). As shown in Fig. 3, we can introduce a saliency detection network[4] (*e.g.*, BAS [56], EGNet [81], ICON [84] or other SOTAs in the recent paper [17]) to obtain the foreground mask and then lock (*e.g.*, bounding box proposal) the domain object. Motivated by the spatial envelope [52] and the block-level strategy [18, 19], we then split the image into different scales (*i.e.*, 1×1, 2×2, $\ldots$, 5×5). These image patches are just like "*kaleido*" patches, and more detailed divisions (*e.g.*, 6×6 or $\mathrm{N}\times\mathrm{N}$ like Pixel-BERT [27]) can be considered according to the difficulty of the specific task. Finally, we obtain 55 kaleido patches from each input image. To generate the embeddings of these patches, we utilize the standard ResNet-50 [25] as our backbone.

3.3. Attention-based Alignment Generator

Attention-based Alignment Generator (**AAG**) aims to find the coarse alignments between text tokens and kaleido patches. As shown in Fig. 4, we directly adopt the famous SAT network [74] as our text generator, which automatically learns to describe the content of images. At the same time, the SAT network generates the attention heat-map for each token, from which we infer the relation between generated tokens and image regions. With the co-occurrence of the generated tokens and the raw text tokens, as well as the overlap area of image regions and kaleido patches, we further build the alignments between raw text tokens and kaleido patches.

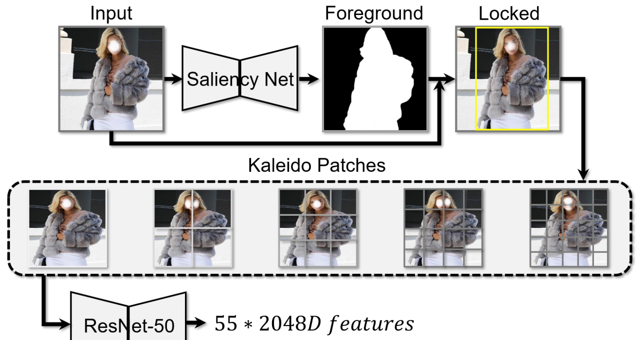

Figure 3: **Illustration of KPG.** See §. 3.2 for details.

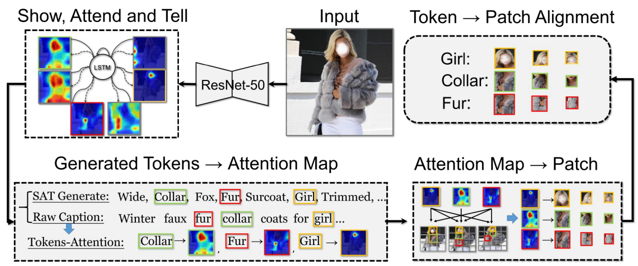

Figure 4: **Procedures of AAG.** See §. 3.3 for details.

3.4. Alignment Guided Masking

The main idea that inspires us to modify the vanilla random masking strategy is that the pre-aligned $\langle$token, patch$\rangle$ pair provides explicit semantic relations between two modalities. This alignment can be used in the pre-training stage, which further forces Kaleido-BERT to *explicitly* explore cross-modality semantic information. As shown in Fig. 2 (Left), unlike the random masking strategy, Alignment Guided Masking (**AGM**) gives high priority to masking the pre-alignment pairs. Meanwhile, for each selected pre-aligned $\langle$token, patch$\rangle$ pair, we randomly mask either the token part or the patch part, which stimulates Kaleido-BERT to learn the missing information in one modality by providing the information of the other. When all pre-alignment pairs are traversed and not-enough tokens or patches are selected, a random masking strategy is adopted to mask the unaligned tokens and patches independently. In this way, we obtain the token and patch masking candidates. AGM strategy works on level-3, level-4, level-5 of kaleido patches. We do not apply this strategy on level-1 & -2 since masking larger patches will increase the

[2] Tensorflow: https://github.com/alibaba/EasyTransfer
[3] Pytorch: https://github.com/huggingface/transformers
[4] For simplicity, we just utilize a very simple UNet-like architecture as our foreground segmentation net.

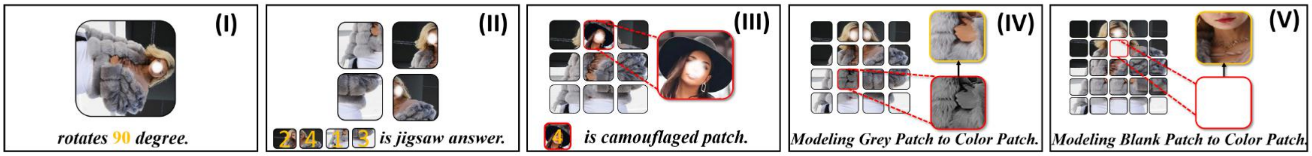

Figure 5: **Aligned Kaleido Patch Modeling (AKPM).** (I) Rotation recognition. (II) Jigsaw puzzles solving. (III) Camouflaged prediction. (IV) Grey-to-color modeling. (V) Blank-to-color modeling. Zoomed-in for a better view. See §. 3.6 for details.

difficulty of modeling. Empirically, we mask one patch in level-3, two patches in level-4, and three patches in level-5.

3.5. Cross-Modality Transformer

We adopt the original BERT [14] as our cross-modality transformer so that our Kaleido-BERT can be easily extended. Spectically, for the text side, we follow Fashion-BERT [21] to encode the order of the token (*i.e.*, generated via WordPieces [72]) position as $0, 1, 2, 3, \ldots, N$. Our final training corpus for each sub-word token is obtained by summing up its embedding with the segment and position embeddings, followed by another layer normalization (LN) layer. For the image side, we encode the position information by re-organizing it as 5D features ($[x_1, x_2, y_1, y_2, w * h]$) for each patch. After that, both patches and location features are fed into a fully-connected (FC) layer in order to project them into the same embedding space. We obtain visual embeddings for each patch by summing up three FC outputs (*i.e.*, FC (seg_id), FC (img_feature), FC (pos_emb))[5] and then passing them through an LN layer.

3.6. Pre-training Kaleido-BERT

To alleviate the VL semantic gap and boost feature representation, we design three pre-training tasks, *i.e.*, Aligned Masked Language Modeling (**AMLM**), Image and Text Matching (**ITM**) and the proposed Aligned Kaleido Patch Modeling (**AKPM**) (which includes five kaleido sub-tasks) to supervise our Kaleido-BERT.

Task #1: AMLM. Derived from our alignment guided masking strategy, we can obtain the mask candidates including both token and patch candidates. When masking indices are determined, we decompose masking words into 10% random words, 10% unchanged, and 80% `[MSK]`. The masked-out token sequence is denoted by $T_i = \{t_1, ...[MSK], ..., t_T\}$, where token t_i is masked out. We feed the hidden output of the last layer of the masked-out token into a classifier over the standard BERT vocabularies. The AMLM goal is to predict the masked words based on the observation of their surrounding tokens and image patches. The objective of the AMLM task is mathematically written as:

$$\mathcal{L}_{AMLM} = \sum CE(t_i, \mathcal{F}(T, K, \theta)_{MSK\_hidden}), \quad (1)$$

where CE denotes the cross-entropy loss. $\mathcal{F}$ is the Kaleido-BERT function. $\mathcal{F}(\cdot)_{MSK\_hidden}$ denotes the hidden output of masked-out tokens. K denotes the masked-out kaleido patch sequence.

Task #2: ITM. The ITM task is transferred by Next Sentence Prediction (NSP) on the vanilla BERT. In this task, `[CLS]` is used to indicate the beginning of the fused representation. The hidden output of `[CLS]` is fed into an FC layer and we use the sigmoid function to predict a score between 0 and 1. The text and image of one positive example are extracted from the same fashion product, while those of one negative sample are randomly extracted from different fashion products. The objective of the ITM task is written as:

$$\mathcal{L}_{ITM} = CE(y_m, \mathcal{F}(T, K, \theta)_{CLS\_hidden}), \quad (2)$$

where y_m denotes the text and image match label.

Task #3: AKPM. The kaleido patch sequence is composed of a collection of kaleido patches as $\{K_1, K_2, ..., K_N\}$, in which N is the kaleido level. As shown in Fig. 5, our AKPM includes N sub-tasks for every kaleido levels, respectively.

Sub-Task #I: Rotation Recognition (RR). Recent works [22, 29] have compared various self-supervised learning strategies concluding that predicting image rotations is among the most effective. Motivated by this, we introduce RR in our pre-training. Specifically, we force the 1×1 patch of the level-1 kaleido to randomly rotate by an angle $\theta \in \{0°, 90°, 180°, 270°\}$. During the training process, we use the angle of the rotated patch as the target label. The hidden output of the K_1 patch is fed into an FC layer followed by softmax function. The final softmax output is used to predict the angle. The objective of the RR task is written as:

$$\mathcal{L}_{RR} = CE(y_r, \mathcal{F}(T, K, \theta)_{K_1\_hidden}), \quad (3)$$

where y_r denotes the rotation angle.

Sub-Task #II: Jigsaw Puzzle Solving (JPS). JPS [29, 51] has been demonstrated to be suitable for self-supervised representation learning. Such a pretext task (also called surrogate task) can mine the spatial relations among image

[5] Similar to 'segment embeddings' in BERT, we conduct a special modality embedding ('T' for text, 'I' for image) to help the model distinguish the different modalities.

patches. Based on this insight, we borrow the notion of jigsaw puzzle to stimulate Kaleido-BERT to learn the potential association from unordered 2×2 patch lists. For simplicity, we treat the JPS problem as a classification of the jigsaw permutations ($4! = 24$ classes). The network architecture is similar to RR. The objective of the JPS task is written as:

$$\mathcal{L}_{JPS} = CE(y_j, \mathcal{F}(T, K, \theta)_{K_2\_hidden}), \tag{4}$$

where y_j denotes the jigsaw permutation.

Sub-Task #III: Camouflage Prediction (CP). To increase the discernment ability of the model, we introduce another camouflage prediction task to judge which patch has been replaced[6]. With the help of image and text clues, this task encourages the training process to observe the diversity among 3×3 patches. We name this task Camouflage Prediction (CP) because its essence is to camouflage one patch then let the model detect it. By pre-training our Kaleido-BERT with CP, the framework achieves a strong capacity to screen out the imparity with varied products. The CP prediction is also treated as a classification problem and its objective is denoted by:

$$\mathcal{L}_{CP} = CE(y_c, \mathcal{F}(T, K, \theta)_{K_3\_hidden}), \tag{5}$$

where y_c denotes the index of a camouflaged patch.

Sub-Task #IV: Grey-to-Color Modeling (G2CM). Different from the masking strategy in existing models, which simply exchanges image embeddings with zero paddings, we introduce a smoother G2CM strategy that greys the image patches. Then we reconstruct the grey patch to a color patch by regression, supervised by KL-divergence, which better caters to self-supervised learning. The objective of G2CM is to minimize the G2CM loss:

$$\mathcal{L}_{G2CM} = \sum KLD(k_{4i}, \mathcal{F}(T, K, \theta)_{K_4\_hidden}), \tag{6}$$

where KLD denotes the KL-divergence, which aims to minimize the distance of the reconstructed distribution to the target distribution and k_{4i} is the masked-out patch(es) of K_4 kaleido patches.

Sub-Task #V: Blank-to-Color Modeling (B2CM). The last sub-task is B2CM. Similar to other pre-training methods that replace image feature embeddings with the same-dimension zeros sequence, we also adopt this kind of patch masking scheme. This strongly tests the learning ability of a model that captures the contextual information. The objective of B2CM is to minimize the B2CM loss:

$$\mathcal{L}_{B2CM} = \sum KLD(k_{5i}, \mathcal{F}(T, K, \theta)_{K_5\_hidden}), \tag{7}$$

where k_{5i} is the masked-out patch of K_5.

[6] The camouflaged patch is the same scale as the patch randomly selected selecting from the other product randomly.

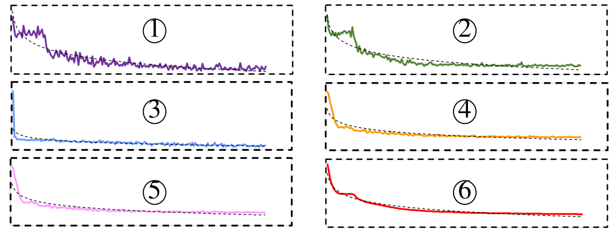

Figure 6: **Evolution of each training loss.** The scores are computed from the validation sets. ①: *Rotation loss.* ②: *Jigsaw loss.* ③: *Camouflage loss.* ④: *Grey-to-Color loss.* ⑤: *Blank-to-Color loss.* ⑥: *Total Loss = AKPM + ITM + AMLM.* This shows that Kaleido-BERT can learn from the kaleido strategy.

All in all, we introduce the aligned kaleido patch modeling to enhance the ability of the model for spatial context structure (*i.e.*, RR and JPS), classification (*i.e.*, CP), and image generation (*i.e.*, G2CM and B2CM). Finally, Kaleido-BERT should minimize the overall loss function as:

$$\begin{aligned}\mathcal{L}_{total} &= \mathcal{L}_{AMLM} + \mathcal{L}_{ITM} + \mathcal{L}_{RR} + \mathcal{L}_{JSP} \\ &+ \mathcal{L}_{CP} + \mathcal{L}_{G2CM} + \mathcal{L}_{B2CM}.\end{aligned} \tag{8}$$

The evaluations of different Kaleido tasks on the validation set are shown in Fig. 6. As can be seen, the losses decay smoothly, which proves that the pre-training process carries on as normal, and the designed tasks can be learned well with Kaleido-BERT.

4. Experiments

We evaluate our Kaleido-BERT on four VL tasks by transferring the pre-trained model to each target task and fine-tuning through end-to-end training.

4.1. Pre-training Settings

Dataset. For a fair comparison, we follow the same settings as the Top-1 FashionBERT [21] and pre-train the proposed Kaleido-BERT on the Fashion-Gen[7] dataset. It contains 67,666 fashion products accompanied with text descriptions. Each product includes one to six images from different angles. Among all the image-text pairs, like [21], we use 260,480 for training, and 35,528 for testing.

Implementation Details. Our Kaleido-BERT has: L=12, H=768, A=12. L is number of stacked Transformer blocks. H denotes the hidden activation, and A means the number of attention heads. We implement our model with Tensorflow and use 8*Tesla V100 for pre-training. The Adam optimizer is applied with a learning rate of $2e-5$ and weight decay $1e-4$. We adopt a warming-up strategy for the first 5K steps.

[7] https://fashion-gen.com/

Table 2: **Retrieval performances on Fashion-Gen dataset.** Here, Sum$\mathcal{R}$=(Rank@1+Rank@5+Rank@10)*100. See §. 4.3 for details.

Tasks		VSE [33]	VSE++ [16]	SCAN [36]	PFAN [71]	ViLBERT [60]	VLBERT [45]	FashionBERT [21]	ImageBERT [55]	OSCAR [42]	Kaleido-BERT *Ours*
1.ITR	Rank@1 ↑	4.010%	4.590%	4.590%	4.290%	20.97%	19.26%	<u>23.96%</u>	22.76%	23.39%	**27.99%** (+4.030%)
	Rank@5 ↑	11.03%	14.99%	16.50%	14.90%	40.49%	39.90%	<u>46.31%</u>	41.89%	44.67%	**60.09%** (+13.78%)
	Rank@10 ↑	22.14%	24.10%	26.60%	24.20%	48.21%	46.05%	52.12%	50.77%	<u>52.55%</u>	**68.37%** (+15.82%)
2.TIR	Rank@1 ↑	4.350%	4.600%	4.300%	6.200%	21.12%	22.63%	<u>26.75%</u>	24.78%	25.10%	**33.88%** (+7.130%)
	Rank@5 ↑	12.76%	16.89%	13.00%	20.79%	37.23%	36.48%	46.48%	45.20%	<u>49.14%</u>	**60.60%** (+11.46%)
	Rank@10 ↑	20.91%	28.99%	22.30%	31.52%	50.11%	48.52%	55.74%	55.90%	<u>56.68%</u>	**68.59%** (+11.91%)
	Sum$\mathcal{R}$ ↑	75.20	94.16	87.29	101.90	218.13	212.84	251.36	241.30	<u>251.53</u>	**319.52**

4.2. Downstream Tasks

We evaluate our model for four downstream VL tasks, including Image-Text Retrieval, Text-Image Retrieval, Category Recognition, and Fashion Captioning. The four tasks strongly cater to industrial applications in the fashion field.

1. Image-Text Retrieval (ITR). Image retrieval is a downstream task that requires the model to distinguish whether a sentence can effectively describe an image. We sample the product images and titles as image-sentences pairs provided by the Fashion-Gen [58] and consider the original product information as positive samples. At the same time, we shuffle the dataset and consider the unmatched image-sentence pairs as negative samples. To increase the difficulty, the positive and negative pairs are selected from the same sub-category, which is hard for PTM to differentiate. We use Rank@1, Rank@5, Rank@10 to evaluate the retrieval performance.

2. Text-Image Retrieval (TIR). The text retrieval task aims to rank product images according to their title. Similar to image retrieval, we use the ground-truth image in the pair as the positive sample and randomly sample 100 unrelated captions from other products in the same sub-category. By predicting the matching score, Rank@1, @5, @10 are used as metrics.

3. Category/SubCategory Recognition (CR & SUB). The category is a vital attribute for describing a product, and is especially useful in many real-life applications. We consider a classification task that judges the category and subcategory of a product, such as {`HOODIES, SWEATERS`}, {`TROUSERS, PANTS`}. We directly use a FC layer after `[CLS]` for these tasks.

4. Fashion Captioning (FC). Image captioning has emerged as an important research topic with a rich literature in computer vision, and the accuracy on FC can evaluate the generation ability of cross-modality models.

4.3. Competing Models

Detailed comparisons for each downstream task are shown in Tab. 2 & Tab. 3. (i) Our Kaleido-BERT achieves significant improvement on nearly all evaluations, which demonstrates its excellent understanding and generation ability in fashion domain. (ii) We observe that the FashionBERT approach outperforms ViLBERT and VLBERT. The main difference between them is that FashionBERT adopts patches as image features, while ViLBERT and VLBERT extract RoIs as features. This indicates that in the fashion domain, the patch method is better for extracting image features. (iii) ImageBERT and Oscar perform better than VLBERT and ViLBERT by adding RoI object classification and RoI tags. These two methods provide more information about the image modality. This, to a certain degree, hints that more image semantic information (*e.g.* image features, image supervision tasks) is required to guide model learning. In our Kaleido-BERT, the kaleido strategy extends from the patch method of FashionBERT [21]. The attention-based alignment masking and the kaleido pre-training task provide more semantic information from the image modality. These factors together explain the superiority of Kaleido-BERT in VL understanding and generation in the fashion domain.

Table 3: **Category Recognition and Fashion Captioning performances on Fashion-Gen dataset.** Here, Sum $\mathcal{CLS}$= (ACC+macro-$\mathcal{F}$)*100 and Sum $\mathcal{CAP}$=Bleu-4+METEOR+ROUGE-L+CIDEr. See §. 4.3 for details.

Tasks		FashionBERT [21]	ImageBERT [55]	OSCAR [42]	Kaleido-BERT *Ours*
3.CR	ACC ↑	91.25%	90.77%	<u>91.79%</u>	**95.07%** (+3.28%)
	macro-$\mathcal{F}$ ↑	0.705	0.699	<u>0.727</u>	**0.714** (−0.013)
3.SUB	ACC ↑	<u>85.27%</u>	80.11%	84.23%	**88.07%** (+2.80%)
	macro-$\mathcal{F}$ ↑	<u>0.620</u>	0.575	0.591	**0.636** (+0.016)
	Sum $\mathcal{CLS}$ ↑	<u>309.02</u>	298.28	307.82	**318.14**
4.FC	Bleu-4 ↑	3.30	-	<u>4.50</u>	**5.70** (+1.2)
	METEOR ↑	9.80	-	<u>10.9</u>	**12.8** (+1.9)
	ROUGE-L ↑	29.7	-	<u>30.1</u>	**32.9** (+2.8)
	CIDEr ↑	30.1	-	<u>30.7</u>	**32.6** (+1.9)
	Sum $\mathcal{CAP}$ ↑	72.9	-	<u>76.2</u>	**84.0**

4.4. Ablation Study

Three main factors may influence the performance of Kaleido-BERT, including Input-level: Kaleido Patch Generator (KPG); Embedding-level: Attention Guided Masking (AGM); and Task-level: Aligned Kaleido Patch Modeling (AKPM). We thus perform three main ablation studies to further analyze these components/strategies of our model. The results are shown in Tab. 4 and Fig. 7.

KPG. Three attempts we have tried to generate our kaleido

Table 4: **Ablation studies of 3 vital pre-training factors.** See §. 4.4 for details.

Metrics		KPG			AGM		AKPM						
		Scale-fixed	Kaleido.	Kaleido.+SOD	Random	AGM	B	B+I	B+I∼II	B+I∼III	B+I∼IV	B+I∼V	B+V
1. Rank@1	↑	24.71	26.73(+8.2%)	27.99(+13.3%)	26.55	27.99(+5.4%)	25.37	25.07(-1.2%)	26.03(+2.6%)	26.88(+6.0%)	26.20(+3.3%)	27.99(+10.3%)	24.62(-2.9%)
1. Rank@5	↑	50.05	54.55(+9.0%)	60.09(+20.1%)	55.13	60.09(+8.9%)	54.97	55.14(+0.3%)	56.31(+2.4%)	58.34(+6.1%)	59.13(+7.6%)	60.09(+9.3%)	53.78(-2.2%)
1. Rank@10	↑	58.93	65.44(+11.0%)	68.37(+16.0%)	64.92	68.37(+5.3%)	62.13	62.90(+1.2%)	63.37(+2.0%)	67.79(+9.1%)	67.99(+9.4%)	68.37(+10.0%)	60.88(-2.0)%
2. Rank@1	↑	30.17	32.19(+6.7%)	33.88(+12.0%)	32.14	33.88(+5.4%)	31.09	30.98(-0.4%)	32.22(+3.6%)	33.17(+6.7%)	33.80(+8.7%)	33.88(+9.0%)	30.77(-1.0%)
2. Rank@5	↑	52.29	58.40(+11.7%)	60.60(+15.9%)	56.99	60.60(+6.3%)	57.35	57.44(+0.2%)	58.73(+2.4%)	58.55(+2.1%)	60.57(+5.6%)	60.60(+5.7%)	55.95(-2.4%)
2. Rank@10	↑	60.82	66.49(+9.3%)	68.59(+12.8%)	63.77	68.59(+7.6%)	64.79	65.65(+1.3%)	64.16(-1.0%)	67.92(+4.8%)	68.41(+5.6%)	68.09(+5.1%)	61.70(-4.8%)
Sum $\mathcal{R}$	↑	**276.97**	**303.80(+9.7%)**	**319.52(+16.2%)**	**299.50**	**319.52(+6.7%)**	**295.70**	**297.18(+0.5%)**	**300.82(+1.7%)**	**312.65(+5.7%)**	**316.10(+6.9%)**	**319.02(+7.9%)**	**287.70(-2.7%)**
3. ACC	↑	93.44%	93.45%(+0.0%)	95.07%(+1.7%)	92.71%	95.07%(+2.5%)	90.94%	90.82%(-0.1%)	91.40%(+0.5%)	93.91%(+3.3%)	94.05%(+3.4%)	95.07%(+4.5%)	88.87(-2.3%)
3. macro-$\mathcal{F}$	↑	0.701	0.705(+0.6%)	0.714(+1.9%)	0.711	0.714(+0.4%)	0.690	0.692(+0.3%)	0.721(+4.5%)	0.713(+3.3%)	0.710(+2.9%)	0.714(+3.5%)	0.701(+1.4%)
4. ACC	↑	86.89%	87.61%(+0.8%)	88.07%(+1.4%)	87.20%	88.07(+1.0%)	81.66%	81.25%(-0.5%)	84.44%(+3.4%)	86.49%(+5.9%)	88.53%(+8.4%)	88.07%(+7.9%)	81.64(+0.0%)
4. macro-$\mathcal{F}$	↑	0.630	0.634(+0.6%)	0.636(+1.0%)	0.633	0.636(+0.5%)	0.558	0.575(+3.0%)	0.596(+6.8%)	0.636(+14.0%)	0.633(+13.4%)	0.636(+14.0%)	0.596(+8.4%)
Sum $\mathcal{CLS}$	↑	**313.43**	**314.96(+0.5%)**	**318.14(+1.5%)**	**314.31**	**318.14(+1.2%)**	**297.40**	**298.77(+0.4%)**	**307.54(+3.4%)**	**315.30(+6.0%)**	**316.88(+6.5%)**	**318.14(+7.0%)**	**300.21(+0.9%)**
5. Bleu-4	↑	4.9	5.2(+6.1%)	5.7(+16.3%)	5.3	5.7(+7.5%)	4.9	5.2(+6.1%)	5.2(+6.1%)	5.1(+4.1%)	5.6(+14.3%)	5.7(+16.3%)	5.3(+8.2%)
5. METEOR	↑	11.0	11.7(+6.4%)	12.8(+16.4%)	11.3	12.8(+13.3%)	11.6	11.6(+0.0%)	11.8(+1.7%)	12.6(+8.6%)	12.8(+10.3%)	12.8(+10.3%)	11.4(-1.7%)
5. ROUGE-L	↑	29.8	31.5(+5.7%)	32.9(+10.4%)	30.3	32.9(+8.6%)	30.4	30.7(+1.0%)	30.8(+1.3%)	31.9(+4.9%)	32.7(+7.6%)	32.9(+8.2%)	30.6(+0.7%)
5. CIDEr	↑	30.9	31.3(+1.3%)	32.6(+5.5%)	31.7	32.6(+2.8%)	31.0	31.5(+1.6%)	31.4(+1.3%)	32.0(+3.2%)	32.3(+4.2%)	32.6(+5.2%)	31.3(+1.0%)
Sum $\mathcal{CAP}$	↑	**76.6**	**79.7(+4.0%)**	**84.0(+9.7%)**	**78.6**	**84.0(+6.9%)**	**77.9**	**79.0(+1.4%)**	**79.2(+1.7%)**	**81.6(+4.7%)**	**83.4(+7.1%)**	**84.0(+7.8%)**	**78.6(+0.9%)**

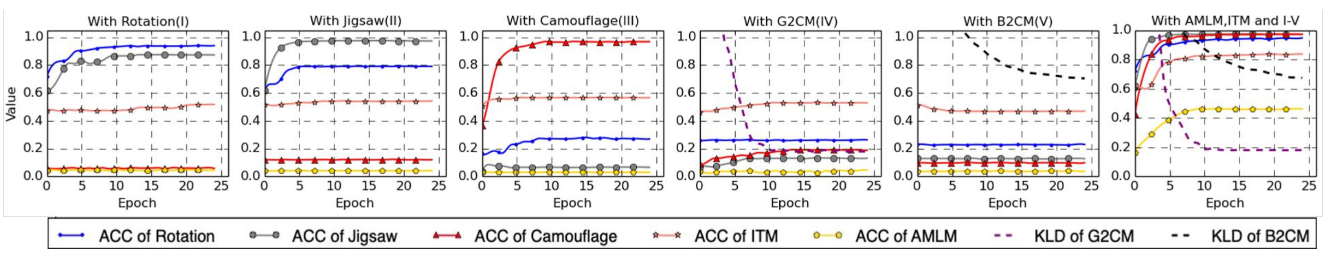

Figure 7: **Single task analysis on our Kaleido-BERT.** ACC = Accuracy, KLD = Kullback-Leibler Divergence. Refer to §. 4.4 for details.

patches. **Scheme-1:** Similar to [21, 34], the first attempt is to split the fashion images with a fixed-scale setting. Training with such patches, we obtain 276.97 Sum $\mathcal{R}$, 313.43 Sum $\mathcal{CLS}$ and 76.6 Sum $\mathcal{CAP}$ scores. **Scheme-2:** We carry out a kaleido (patch-invariant) scheme to generate patches and achieves +9.7%, +0.5% and +4.0% relative improvement on each metric. Compared with scheme-1, scheme-2 is capable of capturing fine-grained representation better. **Scheme-3:** We further introduce the salient object detection (SOD) algorithm [81] to avoid a huge number of patches with blank regions (tabula rasa). We observed 16.2%, +1.5% and +9.7% relative improvement compared with Scheme-1.

AGM. The majority of existing masking methods independently mask visual and text features with a pre-set probability. Such kind masking methods are usually named as random masking. In this experiment, we compare AGM to random masking (Random). Not surprisingly, AGM obtains +6.7%, +1.2%, 6.9% improvements. Compared to random masking, AGM generates more semantic related masking, which benefits our Kaleido-BERT to better understand multi-modality information.

AKPM. To verify the efficiency of the proposed AKPM, we conduct 7 ablation studies (see Fig. 7). The baseline (B) merely consists of the basic ITM and AMLM. Then we add five sub-task to pre-train the model step by step. For example, "B + I ∼ IV" equals to "B + I + II + III + IV". Note that existing models [21] usually use the combination of "ITM + AMLM + B2CM" (B + V) as the pre-training supervision. As shown in Tab. 4, the improvement is limited (+0.9%) in terms of Sum $\mathcal{CLS}$ score, even cause negative effect (-2.7%) in Sum$\mathcal{R}$ metric. Interestingly, we naively replace V (B2CM) with I (RR) will obtains improvement on all downstream tasks (+0.5%, +0.4% and +1.4%, respectively). Gradually, we observed that the performance continue improve when we adding the corresponding subtasks sequentially. Meanwhile the negative affect of V has been alleviated, we argue that V plays the true value when Kaleido-BERT has learned the comprehensive representations of image embeddings from I∼IV.

5. Conclusion and Future Work

We presented a novel pre-trained vision-language understanding architecture **Kaleido-BERT** for fashion-based task. It consists of a kaleido patches generator, attention-based alignment generator, and alignment guided masking strategy. These components are easy to implement and cooperate closely to learn both intra-modal and inter-modal image-text feature embeddings. The designed Kaleido-BERT is much more efficient than existing models, attains the new SOTA performance, and largely boosts the accuracy of many downstream tasks such as Image-Text Retrieval, Category Recognition, and Fashion Captioning.

References

[1] Ziad Al-Halah and Kristen Grauman. From paris to berlin: Discovering fashion style influences around the world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10136–10145, 2020.

[2] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054*, 2019.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015.

[5] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*. Springer, 2020.

[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[8] 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.

[9] 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 *ECCV*, pages 104–120. Springer, 2020.

[10] Meng-Jiun Chiou, Roger Zimmermann, and Jiashi Feng. Rvl-bert: Visual relationship detection with visual-linguistic knowledge from pre-trained representations. *arXiv preprint arXiv:2009.04965*, 2020.

[11] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Weller Adrian. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

[12] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, pages 326–335, 2017.

[13] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. *arXiv preprint arXiv:2006.06666*, 2020.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[15] Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*, 2020.

[16] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[17] Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*, pages 186–202, 2018.

[18] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4548–4557, 2017.

[19] Deng-Ping Fan, ShengChuan Zhang, Yu-Huan Wu, Yun Liu, Ming-Ming Cheng, Bo Ren, Paul L Rosin, and Rongrong Ji. Scoot: A perceptual metric for facial sketches. In *ICCV*, pages 5612–5622, 2019.

[20] 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 *NIPS*, 2020.

[21] 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 *SIGIR*, pages 2251–2260, 2020.

[22] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *ICCV*, pages 6391–6400, 2019.

[23] 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 *CVPR*, pages 6904–6913, 2017.

[24] 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 *CVPR*, pages 13137–13146, 2020.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[26] Wei-Lin Hsiao, Isay Katsman, Chao-Yuan Wu, Devi Parikh, and Kristen Grauman. Fashion++: Minimal edits for outfit improvement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5047–5056, 2019.

[27] 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.

[28] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018.

[29] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *PAMI*, 2020.

[30] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[31] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014.

[32] Aisha Urooj Khan, Amir Mazaheri, Niels da Vitoria Lobo, and Mubarak Shah. Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering. *arXiv preprint arXiv:2010.14095*, 2020.

[33] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[34] Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16-16 words: Transformers for image recognition at scale. In *arXiv preprint arXiv:2010.11929*, 2020.

[35] 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. *IJCV*, 123(1):32–73, 2017.

[36] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018.

[37] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

[38] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020.

[39] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020.

[40] 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.

[41] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does bert with vision look at? In *ACL*, pages 5265–5275, 2020.

[42] 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 *ECCV*, pages 121–137. Springer, 2020.

[43] 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.

[44] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869. Springer, 2016.

[45] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NIPS*, pages 13–23, 2019.

[46] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, pages 10437–10446, 2020.

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

[48] 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. *arXiv preprint arXiv:2004.14973*, 2020.

[49] 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 *ICCV*, pages 2630–2640, 2019.

[50] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *ECCV*, 2019.

[51] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84. Springer, 2016.

[52] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[53] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, pages 1143–1151, 2011.

[54] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015.

[55] 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.

[56] Xuebin Qin, Deng-Ping Fan, Chenyang Huang, Cyril Diagne, Zichen Zhang, Adrià Cabeza Sant'Anna, Albert Suàrez, Martin Jagersand, and Ling Shao. Boundary-aware segmentation network for mobile and web applications. *arXiv preprint arXiv:2101.04704*, 2021.

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

[58] 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.

[59] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018.

[60] 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 *ICLR*, 2020.

[61] 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 *ICLR*, 2020.

[62] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, 2019.

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

[64] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, pages 7464–7473, 2019.

[65] Reuben Tan, Mariya I Vasileva, Kate Saenko, and Bryan A Plummer. Learning similarity conditions without explicit supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10373–10382, 2019.

[66] Trieu H Trinh, Minh-Thang Luong, and Quoc V Le. Selfie: Self-supervised pretraining for image embedding. *arXiv preprint arXiv:1906.02940*, 2019.

[67] Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 390–405, 2018.

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[69] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021.

[70] Yue Wang, Shafiq Joty, Michael R Lyu, Irwin King, Caiming Xiong, and Steven CH Hoi. Vd-bert: A unified vision and dialog transformer with bert. In *EMNLP*, 2020.

[71] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. *arXiv preprint arXiv:1907.09748*, 2019.

[72] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[73] 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. *arXiv preprint arXiv:2003.01473*, 2020.

[74] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.

[75] Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. Fashion captioning: Towards generating accurate descriptions with semantic rewards. *arXiv preprint arXiv:2008.02693*, 2020.

[76] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NIPS*, pages 5753–5763, 2019.

[77] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, pages 21–29, 2016.

[78] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020.

[79] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731, 2019.

[80] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*, 2019.

[81] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *ICCV*, pages 8779–8788, 2019.

[82] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, pages 13041–13049, 2020.

[83] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, pages 8746–8755, 2020.

[84] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *arXiv preprint arXiv:2101.07663*, 2021.