Machine Intelligence Research
www.mi-research.net

20(3), June 2023, 421-434
DOI: 10.1007/s11633-022-1394-4

Masked Vision-language Transformer in Fashion

Ge-Peng Ji[1†] Mingchen Zhuge[1†] Dehong Gao[1] Deng-Ping Fan[2*]
Christos Sakaridis[2] Luc Van Gool[2]

[1] International Core Business Unit, Alibaba Group, Hangzhou 310051, China

[2] Computer Vision Lab, ETH Zürich, Zürich 8092, Switzerland

Abstract: We present a masked vision-language transformer (MVLT) for fashion-specific multi-modal representation. Technically, we simply utilize the vision transformer architecture for replacing the bidirectional encoder representations from Transformers (BERT) in the pre-training model, making MVLT the first end-to-end framework for the fashion domain. Besides, we designed masked image reconstruction (MIR) for a fine-grained understanding of fashion. MVLT is an extensible and convenient architecture that admits raw multi-modal inputs without extra pre-processing models (e.g., ResNet), implicitly modeling the vision-language alignments. More importantly, MVLT can easily generalize to various matching and generative tasks. Experimental results show obvious improvements in retrieval (rank@5: 17%) and recognition (accuracy: 3%) tasks over the Fashion-Gen 2018 winner, Kaleido-BERT. The code is available at https://github.com/GewelsJI/MVLT.

Keywords: Vision-language, masked image reconstruction, transformer, fashion, e-commercial.

Citation: G. P. Ji, M. Zhuge, D. Gao, D. P. Fan, C. Sakaridis, L. V. Gool. Masked vision-language transformer in fashion. *Machine Intelligence Research*, vol.20, no.3, pp.421–434, 2023. http://doi.org/10.1007/s11633-022-1394-4

1 Introduction

The emergence of transformer is drawing enormous attention from the academic community, facilitating the advancement of computer vision (CV)[1, 2] and natural language processing (NLP)[3, 4]. Benefiting from the robustness of transformers, researchers also contribute to the vision-language (VL) field[5–9] with zeal. To better utilize the pre-trained models in CV and NLP, existing general VL models are mainly based on the BERT model[10] or adopt the well-pretrained vision extractors[11, 12] or both. However, general VL methods[13–15] still struggle when applied to the fashion domain in e-commerce because they suffer from two main issues: **1) Insufficient Granularity.** Unlike the general objects with complex backgrounds, only focusing on coarse-grained semantics is insufficient for a fashion product[16–18], as it would lead the network to generate sub-optimal results. Contrarily, the fashion-oriented framework requires more fine-grained representations, such as a suit with different materials (e.g., wool, linen, and cotton) or collars (e.g., band, camp, and Windsor). **2) Bad Transferability.** The pre-extracted visual features are not discriminative for fashion-oriented tasks, restricting the cross-modal representations.

To address the above issues, we present a novel VL framework termed masked vision-language transformer (MVLT). Specifically, we introduce a generative task, masked image reconstruction (MIR), for the fashion-based VL framework. Compared to previous pre-training tasks, such as masked image modeling (regression task) or masked image classification (classification task), MIR enables the network to learn more fine-grained representations via pixel-level visual knowledge (see Fig. 1). Further, inspired by pyramid vision transformer (PVT)[21], we utilize a pyramid architecture for our VL transformer. Then, we introduce the MIR task. These two improvements significantly enhance the ability to adapt to fashion-specific understanding and generative tasks and can conduct in an end-to-end manner. To this end, MVLT can directly process the raw multi-modal inputs in dense formats (i.e., linguistic tokens and visual patches) without extra (e.g., ResNet) pre-processing models[22, 23]. Our main contributions are summarized as follows:

1) We introduce a novel MIR task, which is the first real pixel-level generative strategy utilized in VL pre-training.

2) Based on the MIR task, we present an end-to-end VL framework, called MVLT, for the fashion domain, greatly promoting the transferability to the downstream tasks and large-scale web applications.

3) Extensive experiments show that MVLT significantly outperforms the state-of-the-art models in matching and generative tasks.

Research Article
Manuscript received on May 24, 2022; accepted on October 14, 2022; published online on February 27, 2023
Recommended by Associate Editor Chun-Hua Shen
Colored figures are available in the online version at https://link.springer.com/journal/11633

*Corresponding author. Work was done while Ge-Peng Ji was a research intern at Alibaba Group.

†These aughors contribute equally to this paper

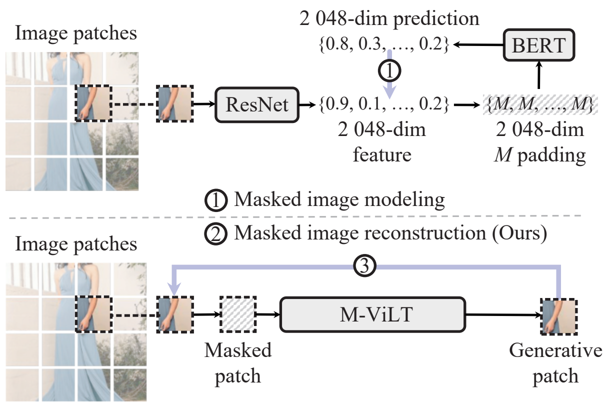


Fig. 1 Different visual reconstruction tasks for VL pretraining[19, 20] utilize masked image modeling (top) with the random masking strategy (i.e., to use M padding to replace raw vectors), which reconstructs pre-extracted visual semantics (i.e., probabilities) at the feature-level. We introduce a generative task named masked image reconstruction (bottom), which directly reconstructs image patches at the pixel level.

2 Background

In recent years, BERT-based pre-training models have been widely investigated in VL tasks. Many previous attempts, such as LXMERT[24], VL-BERT[25], and Fashion-BERT[19], were successful in a wide range of downstream applications. Experiments and discussions show that BERT is a powerful method for learning multi-modal representations, outperforming several previous CNN-based[26] or LSTM-based[27, 28] approaches. Compared to previous studies, this paper aims to develop a more efficient self-supervised objective that can be easily implemented in pre-training and provides better representations for real-world applications. Thus, we review research on masked learning strategies and end-to-end multi-modal schemes that inspired us the most.

2.1 Masked learning strategies

Masked modeling is a vital self-supervised task in BERT[10] and initially demonstrates outstanding abilities in natural language processing. Researchers have replicated its strength in language models because of its utility in multi-modal and vision tasks. Most VL works[14, 25, 29] transfer masked modeling into visual tokens and use a regression task to construct the token feature from nonsense-replace or a classification task to predict the token's attribute. To reduce the difficulty in learning, Kaleido-BERT[2] optimizes masked modeling by employing a Kaleido strategy that facilitates coherent learning for multi-grained semantics. Although this work improves the performance of VL-related tasks in fashion indeed, we argue that the token-patch pre-alignment scheme by using an auxiliary tool[30, 31] is still complex and impedes the application to practical settings. Another work[32] introduces the masked language and image modeling (MLIM) approach that strengthens masked image modeling with an image reconstruction task, which shares a similar idea to ours. However, our experiments showed that requiring a model to reconstruct the entire image without any reminder is too difficult. Recently, BEiT[33] and MAE[34] utilized a BERT-style pre-training as part of the visual learner, and they discovered that models are effective at learning semantics with such a scheme. These two works strengthen our conviction that converting the original masked image modeling (i.e., a regression task) to a masked image reconstruction task is possible. However, our primary goal is to design a generative pretext task that makes the multi-modal modeling in VL pre-training easier while eliminating the need for using prior knowledge. It will be extremely helpful in our practical application setting with billion-level data.

2.2 End-to-end multi-modal schemes

Pixel-BERT[35] is the first method to consider end-to-end pre-training. It employs 2×2 max-pooling layers to reduce the spatial dimension of image features, with each image being downsampled 64 times. Although this work sets a precedent for end-to-end training, such a coarse and rigid method cannot work well in practical settings because it is simply combined with a ResNet[11] as part of joint pre-training without considering the loss in speed and performance. Recently, VX2TEXT[36] proposed to convert all modalities into a language space and perform end-to-end pre-training using a relaxation scheme. Though it is exciting to translate all the modalities into a unified latent space, it ignores that the usage of data extracted by pre-trained methods as input to the model cannot be regarded as an end-to-end framework. According to the timeline, ViLT[37] is the first method that indeed investigates an end-to-end framework via replacing region-based or grid-based features with patch-based projections. However, without other designs, it cannot obtain competitive performance since it is just a vanilla extension of ViT[1]. Grid-VLP[38] is similar to ViLT, but it takes a further step by demonstrating that using a pre-trained CNN network as the visual backbone can improve performance on downstream tasks. SOHO[39] takes the entire image as input and creates a visual dictionary to affine the local region. However, this method does not fit fashion-specific applications due to the lack of reliable alignment information. As a result, the vision dictionary may merely learn the location of the background or foreground rather than complex semantics. FashionVLP[40] uses a feedback strategy to achieve better retrieval performance. In practice, they use the well-pretrained knowledge extracted from ResNet and then model the whole, cropped, and landmark representations. Besides, they adopt Faster-RCNN as an object detector for popping out RoI candidates. Besides, some works are designed for end-to-end pre-training[41−43], but they are used for specific tasks and are not directly applicable to our research.

Despite existing methods employing different approaches to construct an end-to-end scheme, solutions that forgo pre-trained methods (e.g., ResNet, BERT) and use raw data (i.e., text, image) as inputs remain underexplored and are needed urgently in multi-modal applications.

Remark. As shown in Fig. 2, similar to the existing two fashion-based approaches, i.e., FashionBERT (a) and Kaleido-BERT (b), the proposed MVLT (c) is also a patch-based VL learner, which extends the pyramid vision transformer[21] to an architecture that adaptively extracts hierarchical representations for fashion cross-modal tasks. It is the first model that solves the end-to-end problem of VL pre-training in fashion, which allows us to simplify the implementation of our MVLT in the fashion industry using a twin-tower architecture[44].

3 Masked vision-language transformer

Our goal is to build an end-to-end VL framework for the fashion domain. The overall pipeline of our MVLT is depicted in Fig. 3. Like PVT, our architecture inherits four stages' properties and generates different-sized features. Two keys of the proposed architecture are the multi-modal encoder (Section. 3.1) and the pre-training objectives (Section. 3.2).

3.1 Multi-modal encoder

As shown in Fig. 3, MVLT accepts visual and verbal inputs. On the language side, we first tokenize the caption of a fashion product and use the specific token [MASK]

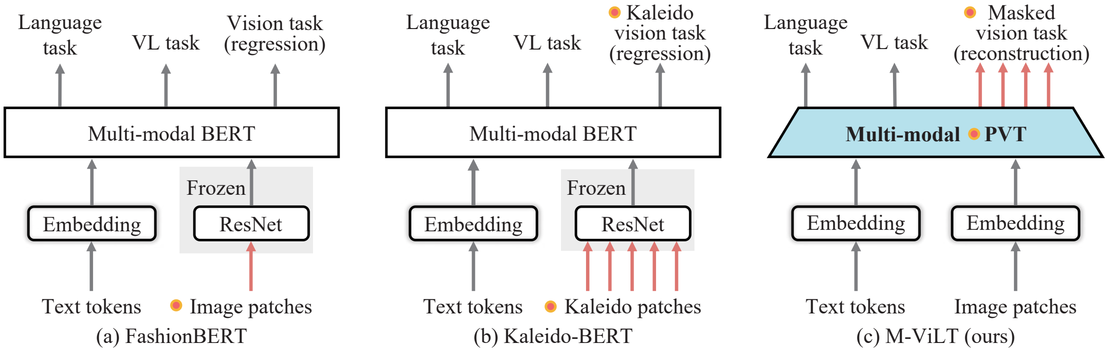


Fig. 2 Comparison of MVLT to cutting-edge fashion-oriented VL frameworks. FashionBERT (a) utilizes a language-based encoder (i.e., BERT) to extract VL representations with single-scale visual input (i.e., image patches). Kaleido-BERT (b) extends it with two upgrades: adds five fixed-scale inputs (i.e., Kaleido patches) to acquire hierarchical visual features and designs Kaleido vision tasks to fully learn VL representations. However, the visual embedding of these models is frozen (i.e., without parameter updating); thus, a lack of domain-specific visual knowledge severely hinders their transferability. Differently, our MVLT (c) adaptively learns hierarchical features by introducing masked vision tasks in an end-to-end framework, significantly boosting the VL-related understanding and generation.

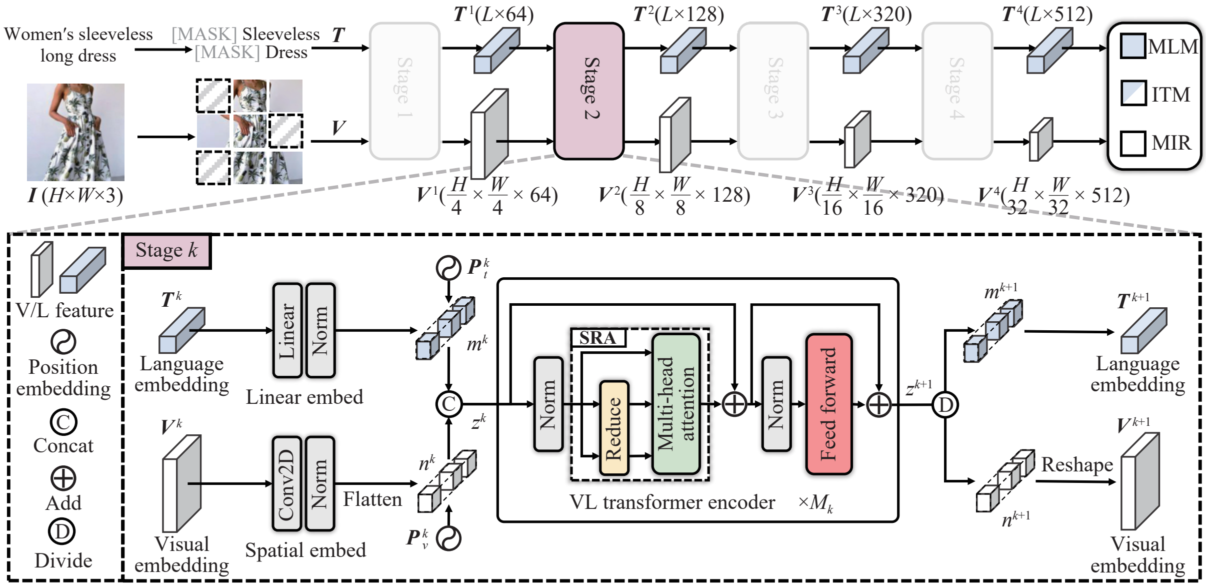


Fig. 3 Pipeline of our MVLT framework. Our overall architecture consists of four stages containing language and visual embeddings and multiple transformer encoders ($\times M_k$). Introducing the masking strategy for three sub-tasks, i.e., masked image reconstruction (MIR), image-text matching (ITM), and masked language modeling (MLM), our MVLT can be trained in an end-to-end manner. More details can be found in Section 3.

to randomly mask out the caption tokens with the masking ratio[1] r_l. Following the masking procedure, we obtain a sequence of word tokens. Then, we insert a specific [CLS] token at the head of this sequence. Besides, we pad the sequence to a unified length L using the [PAD] token if the length is shorter than 128. This procedure generates the language input ids $\boldsymbol{T} \in \mathbf{R}^{L} = \langle t_1; \cdots; t_L \rangle$. On the vision side, we treat $\boldsymbol{I} \in \mathbf{R}^{H\times W\times 3}$ as visual input, where H and W denote the height and width of the given input. This input is sliced into multiple grid-like patches $\boldsymbol{V} \in \mathbf{R}^{N\times P\times P\times 3} = \langle v_1; \cdots; v_N \rangle$, where $N = \frac{HW}{P^2}$ is the total number of patches, and P denotes the patch size. Similarly, the split patches are masked out with mask ratio r_v. We provide more details about the above masking strategy for the language and vision parts in Section 3.2.

The above multi-modal inputs are embedded and fed into the consequent four VL interaction stages (i.e., $k \in \{1, 2, 3, 4\}$). In the first stage, we generate the vision and language embeddings, $\boldsymbol{T}^1$ and $\boldsymbol{V}^1$, respectively, via the given inputs ($\boldsymbol{T}$ and $\boldsymbol{V}$). Regarding the subsequent stages, we consider only the k-th stage to have concise illustrations. As shown in the bottom part of Fig. 3, we first embed the language embedding $\boldsymbol{T}^k \in \mathbf{R}^{L\times D_k}$ into the language hidden feature $m^k \in \mathbf{R}^{L\times D_{k+1}}$, which is formulated as

$$m^k = \boldsymbol{T}^k * \boldsymbol{W}_t^k + \boldsymbol{P}_t^k \tag{1}$$

where $\boldsymbol{W}_t^k \in \mathbf{R}^{D_k\times D_{k+1}}$ and $\boldsymbol{P}_t^k \in \mathbf{R}^{L\times D_{k+1}}$ are the learnable linear embedding and position embedding matrices, D_k is the size of the hidden feature embedding, $*$ denotes the matrix multiplication.

The visual embeddings are $\boldsymbol{V}^k \in \mathbf{R}^{\frac{H}{R_k}\times\frac{W}{R_k}\times D_k}$, where R_k denotes the spatial reduction factor of visual embedding. To acquire pyramid visual features, $\boldsymbol{V}^k$ is then embedded and flattened into the visual hidden feature $n^k \in \mathbf{R}^{(HW/R_{k+1}^2)\times D_{k+1}}$ via a two-dimensional projection (i.e., Conv2D block). In particular, this projection enforces the network to reduce the equivalent spatial dimension from $\mathbf{R}^{HW/R_k^2}$ to $\mathbf{R}^{HW/R_{k+1}^2}$ by utilizing the convolutional kernel $\boldsymbol{W}_v^k \in \mathbf{R}^{D_k\times K_k\times K_k\times D_{k+1}}$ with kernel size K_k and stride length S_k. This could be formulated as follows:

$$n^k = \boldsymbol{Flatten}(\boldsymbol{V}^k * \mathbf{W}_v^k) + \boldsymbol{P}_v^k \tag{2}$$

where $\boldsymbol{P}_v^k \in \mathbf{R}^{N\times D_{k+1}}$ denotes the position embedding matrix, $*$ denotes the matrix multiplication. We then concatenate these two VL hidden features $z^k = \langle m^k; n^k \rangle$ and feed them into multiple (M_k) VL transformer encoders. Each encoder contains the multi-head self-attention layer with spatial reduction (i.e., reduce block), multi-layer perceptron, and layer normalization. Finally, we obtain the encoded multi-modal feature $z^{k+1} = \langle m^{k+1}; n^{k+1} \rangle$ and divide it into a language part $\boldsymbol{T}^{k+1} = m^{k+1}$ and a visual part $\boldsymbol{V}^{k+1} = \boldsymbol{Reshape}(n^{k+1})$, where the $\boldsymbol{Reshape}(\cdot)$ operation consists in recovering the spatial dimension of the given feature.

After four VL interaction stages, we generate the four text embeddings $\{\boldsymbol{T}^k\}_{k=1}^4$ and four pyramid vision embeddings $\{\boldsymbol{V}^k\}_{k=1}^4$, respectively. Table 1 presents more detailed hyperparameter settings of our method.

Table 1 Hyperparameter of our multi-modal encoders

Hyperparameter	$k=1$	$k=2$	$k=3$	$k=4$
Layer number M_k	2	2	2	2
Hidden size D_k	64	128	320	512
Reduction size R_k	4	8	16	32
Kernel size K_k	4	2	2	2
Stride length S_k	4	2	2	2

3.2 Pre-training objectives

To acquire discriminative multi-modal representations, we adopt three pre-training tasks to establish the inter- and intra-relationships between the most primitive VL modalities, including vision (masked image reconstruction, MIR), language (i.e., masked language modeling, MLM), and VL (image-text matching, ITM) modalities.

Objective 1: Masked image reconstruction (MIR). As for the general domain, models are enough to learn the coarse-grained semantics from the patch-based or region-based objectives and achieve satisfactory results. However, the fashion-specific models require more fine-grained representations, such as a suit with different materials (e.g., wool) or collars (e.g., Windsor), which needs a pixel-to-pixel vision pre-training objective. Inspired by masked language modeling[10], we attempt to build pixel-to-pixel relationships from the perspective of generative tasks, which promote the scalability of visual representations. We designed the masked image reconstruction to accomplish this idea. To help our model learn better by MIR, we utilize the pyramid characteristic of the PVT architecture[21] to design a flexible masking strategy. Unlike the ViT-based method (a) in Fig. 4, PVT-based architecture (b) masks out the input image according to the masking unit matrix that contains small-grained patches. Given the patch sequence $\boldsymbol{V} = \{v_n\}_{n=1}^N \in \mathbf{R}^{N\times P\times P\times 3}$, the masked-out sequence $\boldsymbol{V}_{\setminus\Phi}$ is defined as

$$\begin{aligned}&\boldsymbol{V}_{\setminus\Phi} = \mathcal{F}_M(\{\boldsymbol{M}(q;\alpha;\Phi)\}_q^Q, \{v_n\}_{n=1}^N) = \\ &\begin{cases}[\texttt{ZERO}], & \text{if} \quad \boldsymbol{M}(q;\alpha;\Phi) = 1\\ v_n, & \text{if} \quad \boldsymbol{M}(q;\alpha;\Phi) = 0\end{cases}\end{aligned} \tag{3}$$

where $\mathcal{F}_M(\cdot;\cdot)$ represents a function (or procedure) of our masking strategy, q is the randomly selected area of the masking unit, and [ZERO] means that we use a pixel value

[1] We follow the default setting in BERT[10].

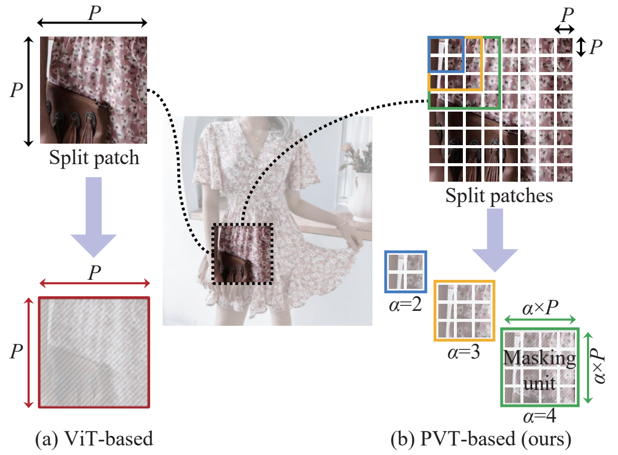


Fig. 4 PVT-based architectures offer more options for designing the masking strategy. The vanilla ViT-based method (a) only selects a fixed-scale patch to mask, i.e., P^2. However, the PVT-based method (b) is more versatile because it combines more fine-grained patches as a basic masking unit, i.e., $(\alpha \times P)^2$, where $\alpha \in \{1, 2, \cdots, 8\}$. These masked patches are not overlapped with each other. This characteristic provides a flexible way to learn the suitable semantics by using different values for α. Notably, we adopt a fixed scale factor of masking units in an individual experiment.

of zero[2] to fill the selected areas. The masking units $\{\boldsymbol{M}(q;\alpha;\Phi)\}_{q=1}^{Q}$ are derived from the indicator function:

$$\boldsymbol{M}(q;\alpha;\Phi) = \mathbf{1}(q) = \begin{cases} 1, & \text{if} \quad q \in \Phi \\ 0, & \text{if} \quad q \notin \Phi \end{cases} \tag{4}$$

where each value in a set of integers Φ is randomly selected from range $[1, Q]$ with ratio r_v. $Q = \frac{H \times W}{(\alpha \times P)^2}$ is the total number of masking units. For instance, in Fig. 4(b), we can define α from 1 to 8. In our default settings, we set $\alpha = 4$ to capture more fine-grained semantics[3].

Since the smooth-$\ell1$ loss is less sensitive to the outliers, we use it as the pre-training objective to reconstruct the whole image via the masked-out sequence $\boldsymbol{V}_{\backslash\Phi}$. It is defined as

$$\mathcal{L}_{\text{MIR}} = \begin{cases} 0.5 \times (\boldsymbol{I}'_{(x,y)} - \boldsymbol{I}_{(x,y)})^2, & \text{if } \boldsymbol{I}'_{(x,y)} - \boldsymbol{I}_{(x,y)} < 1 \\ \mid \boldsymbol{I}'_{(x,y)} - \boldsymbol{I}_{(x,y)} \mid -0.5, & \text{otherwise} \end{cases} \tag{5}$$

where $\boldsymbol{I}'_{(x,y)}$ and $\boldsymbol{I}_{(x,y)}$ denote the pixel at coordinate (x, y) in the reconstructed image $\boldsymbol{I}'$ and the input image $\boldsymbol{I}$, respectively. $\boldsymbol{I}' = \mathcal{F}_{\text{MIR}}(\boldsymbol{V}_{\backslash\Phi}; \boldsymbol{W}_{\text{MIR}})$ is parameterized by learnable weights $\boldsymbol{W}_{\text{MIR}}$. Function $\mathcal{F}_{\text{MIR}}(\cdot;\, \boldsymbol{W}_{\text{MIR}})$ denotes a standard four-level U-Net[45] decoder, which admits four pyramidal vision embeddings $\{\boldsymbol{V}^k\}_{k=1}^{4}$ as inputs.

Objective 2: Image-text matching (ITM). The appended classification embedding in the last language embedding, $\boldsymbol{T}^4$, is used to couple the representations from VL modalities. We utilize the function $\mathcal{F}_{\text{ITM}}(\cdot; \boldsymbol{W}_{\text{ITM}})$ to denote a fully connected (FC) and softmax layers, parameterized by the weights $\boldsymbol{W}_{\text{ITM}}$. $\mathcal{F}_{\text{ITM}}$ outputs a two-class probability vector $\boldsymbol{p}_{\text{ITM}} = \mathcal{F}_{\text{ITM}}(\langle \boldsymbol{T}, \boldsymbol{V} \rangle; \boldsymbol{W}_{\text{ITM}})$, representing whether the input fashion image and caption match (i.e., positive pair) or not (i.e., negative pair). The positive pairs are selected from the same fashion product category, whereas the negative pairs are chosen randomly from different entries. The binary cross-entropy loss function finally constrains this task:

$$\begin{aligned} \mathcal{L}_{\text{ITM}} = &- E_{\langle \boldsymbol{T},\boldsymbol{V} \rangle}[\boldsymbol{y}_{\text{ITM}} \log(\boldsymbol{p}_{\text{ITM}}) + \\ &(1 - \boldsymbol{y}_{\text{ITM}}) \log(1 - \boldsymbol{p}_{\text{ITM}})] \end{aligned} \tag{6}$$

where $\boldsymbol{y}_{\text{ITM}}$ denotes the ground-truth label, i.e., 1 for matched pairs and 0 for unmatched pairs.

Objective 3: Masked language modeling (MLM). Following [46], we randomly use the specific token [MASK] to replace the original text tokens. The target of the MLM is to predict the text content for the masked tokens using the unmasked tokens and patches. Given a tokenized sequence $\boldsymbol{T} = \{t_1, \cdots, t_L\}$, the masked-out sequence is denoted by $\boldsymbol{T}_{\backslash i} = \{t_1, \cdots, [\texttt{MASK}]_i, \cdots, t_L\}$. We use the cross-entropy loss to model this objective:

$$\mathcal{L}_{\text{MLM}} = -E_{\boldsymbol{T}}[\log(\boldsymbol{p}_{\text{MLM}})] \tag{7}$$

where $\boldsymbol{p}_{\text{MLM}} = \mathcal{F}_{\text{MLM}}(\boldsymbol{T}_{\backslash i}; \boldsymbol{W}_{\text{MLM}})$ denotes the predicted probability for each masked-out token $[\texttt{MASK}]_i$ using $\boldsymbol{T}_{\backslash i}$. The function $\mathcal{F}_{\text{MLM}}(\cdot; \boldsymbol{W}_{\text{MLM}})$ represents the parameters $\boldsymbol{W}_{\text{MLM}}$ of a classifier. The final pre-training objective of the proposed MVLT is a combination of the three objectives:

$$\mathcal{L}_{\text{total}} = w_1 \times \mathcal{L}_{\text{MIR}} + w_2 \times \mathcal{L}_{\text{ITM}} + w_3 \times \mathcal{L}_{\text{MLM}}. \tag{8}$$

3.3 Downstream tasks

For a fair comparison, we follow the same training/inference protocols as in [19, 20] and adopt the Fashion-Gen 2018[47] benchmark as the base of our experiments. This dataset contains 67 666 fashion products (i.e., 60 147 entries for training and 7 519 entries for testing) and their associated product descriptions. Each product corresponds to an image set (including $1-6$ samples) at various viewing angles. As a result, we utilize 260 480 and 35 528 image-text pairs as training and testing partitions, respectively. For a fair comparison, we tested MVLT and compared models on Fashion-Gen using the following four fashion-related VL downstream tasks.

[2] In fact, we set $[\texttt{ZERO}] = 10^{-6}$ to bring better optimization stability and less pattern degradation.

[3] The vanilla masking strategy in Fig. 4(a) with $P = 32$ becomes a special case of our masking strategy in Fig. 4(b) when $\alpha = 8, P = 4$.

Task 1: Text-image retrieval (TIR). The TIR task requires the model to find a text with the highest similarity value with different query images. In particular, we take a product title and its corresponding image as a positive image-text pair, while the negative pairs are randomly selected from a pool of mismatched images. To increase our experiment's difficulty, we constrain a set of image-text candidates (i.e., a positive pair and 100 negative pairs) in the same sub-category, making them as similar as possible.

Task 2: Image-text retrieval (ITR). As the reverse process of the TIR task, the ITR task aims to retrieve a matching image given a sequence of text entries of fashion description, where these bidirectional retrieval tasks (i.e., TIR and ITR) become prominent members of cross-modal research. Similar to the above selection strategy in the TIR, we prepare a set of candidate image-text pairs, including a positive pair and 100 negative pairs from the same sub-category. We evaluate the zero-shot learning ability of our MVLT without further fine-tuning for these two retrieval tasks. We utilize three accuracy metrics (i.e., $\mathcal{R}$@1, $\mathcal{R}$@5, and $\mathcal{R}$@10) for the evaluation by ranking a series of predicted probabilities.

Task 3: Category recognition (M-CR and S-CR). This task has two parts: main-category recognition (M-CR) and sub-category recognition (S-CR). These tasks are the fundamental role of practical e-commerce applications that offer the specific category of the queried product. We expect that the model should possess the ability to recognize differences under different granularity levels: 48 main-categories and 122 sub-categories, such as $\{\mathrm{M-CR}=\texttt{SWEATERS}, \mathrm{S-CR}=\texttt{CREWNECKS}\}$. After the class embedding in the last language embedding $\boldsymbol{T}^4$, we add two independent FC layers to generate the final probabilities for two different recognition tasks. This procedure requires additional fine-tuning with recognition labels. We utilize two recognition-related metrics to evaluate performance: accuracy ($\mathcal{A}$) and macro F-measure (macro-$\mathcal{F}$).

Task 4: Masked image generation (MIG). The MIG task can be viewed as a pixel-wise reconstruction task. Each patch in the image is randomly masked with the probability r_v (refer to the pre-training task MIR in Section 3.2). Then, we ask the model to recreate the whole image using the uncovered areas as visual clues.

4 Experiments

This section will detail our experiment to determine the factors leading to the success of the proposed MVLT.

4.1 Settings

This part provides the hyperparameter settings for our training procedure: **1) Pre-training.** We utilize PyTorch to implement our method, which is accelerated by 8 Tesla V100 GPUs. We adopt an AdamW optimizer with a momentum value of 0.9, a mini-batch size of 1 200 (i.e., 150 per GPU), and a weight decay of 10^{-4}. To avoid over-fitting, we initialize MVLT on ImageNet pre-trained weights[21]. The learning rate is initially set to 2.5×10^{-3} and is changed using a cosine learning schedule. For the visual side, the input image is resized to $H=W=256$ and split into multiple sub-patches with a size of $P=4$. For the language side, all the product captions are tokenized and padded to tokens with a unified length of $L=128$, including classification, caption, and padding tokens. The mask probabilities for vision and language are set to $r_v=0.5$ and $r_l=0.15$, respectively. We empirically set weighting factors $\{w_1=10, w_2=1, w_3=1\}$ to balance the orders of magnitude of different loss values. **2) Fine-tuning.** We transfer the pre-trained VL representation to each downstream application via fine-tuning in an end-to-end manner, whose settings are consistent with the pre-training process.

4.2 Results

As described in Section 3.3, we provide the details of four downstream fashion-related tasks. Experimental results show that our MVLT outperforms all competitors, including VSE[48], VSE++[49], SCAN[26], PFAN[50], ViLBERT[14], ImageBERT[13], FashionBERT[19], VL-BERT[25], OSCAR[29], and Kaleido-BERT[20], which demonstrate the superiority for handling the VL understanding and generation tasks.

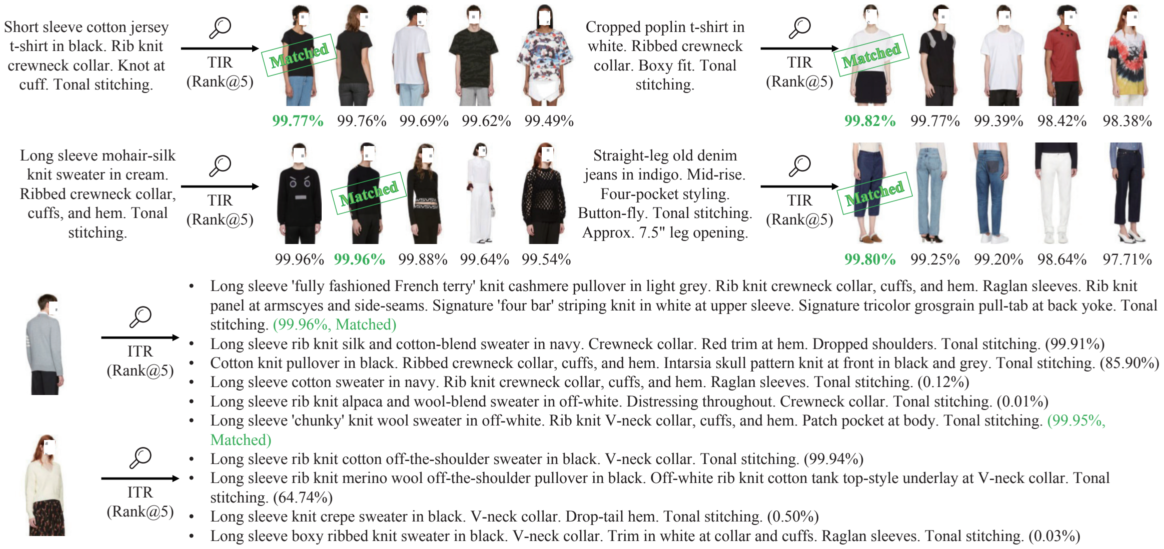
TIR and ITR. As shown in Table 2, our MVLT surpasses the best method (i.e., Kaleido-BERT-CVPR$_{21}$) on the TIR task by margins of +17.40% and +20.91% across the $\mathcal{R}$@5 and $\mathcal{R}$@10. As for ITR, our method delivers more competitive results, with improvements of +17.11% and +22.73% on the $\mathcal{R}$@5 and $\mathcal{R}$@10 metrics, respectively. In any case, these results strongly support that our model is powerful enough to match vision and language. They also show how 1) MIR and 2) end-to-end pre-training are useful in fashion. We believe that MVLT would set a precedent in many industrial applications because it is a simple, cost-effective, and powerful architecture. Besides, we present the visualization results of these two retrieval tasks in Fig. 5.

M-CR and S-CR. Compared with BERT-based architectures[13, 19, 20, 29], we also achieve top-1 performances in these two tasks, demonstrating our method has an excellent VL understanding capability. Moreover, compared with the best method Kaleido-BERT, our architecture improves by 0.193 in the macro-$\mathcal{F}$ metric for the S-CR task. In addition, the mean improvements in terms of the Sum$\mathcal{C}$ metric (i.e., M-CR: +21.39 and S-CR: +24.80) are very significant. Since this metric is very sensitive to data distribution, it demonstrates that MVLT has super-strong robustness. We also present the recognition results of M-CR and S-CR in Fig. 6.

MIG. As shown in Fig. 7, we showcase reconstructed

Table 2 Retrieval (i.e., TIR and ITR) and recognition (i.e., M-CR and S-CR) performances on the Fashion-Gen dataset. ↑ means the larger, the better. Here, Sum$\mathcal{R}$=($\mathcal{R}$@1+$\mathcal{R}$@5+$\mathcal{R}$@10) ×100 and Sum$\mathcal{C}$=($\mathcal{A}$ + macro-$\mathcal{F}$) × 100. "N/A" means the score is not available. "Diff" means the numerical difference between the performance of the second-ranked competitor and our MVLT.

			VSE	**VSE++**	**SCAN**	**PFAN**	**ViLBERT**	**ImageBERT**	**FashionBERT**	**VL-BERT**	**OSCAR**	**Kaleido-BERT**	**MVLT**	
Task	Metric		arXiv$_{14}$	BMVC$_{18}$	ECCV$_{18}$	arXiv$_{19}$	NeurIPS$_{19}$	arXiv$_{20}$	SIGIR$_{20}$	ICLR$_{20}$	ECCV$_{20}$	CVPR$_{21}$	**OUR$_{22}$**	Diff
TIR	$\mathcal{R}$@1	↑	4.350%	4.600%	4.300%	6.200%	21.12%	24.78%	26.75%	22.63%	25.10%	33.88%	**34.60%**	+0.72%
	$\mathcal{R}$@5	↑	12.76%	16.89%	13.00%	20.79%	37.23%	45.20%	46.48%	36.48%	49.14%	60.60%	**78.00%**	+17.40%
	$\mathcal{R}$@10	↑	20.91%	28.99%	22.30%	31.52%	50.11%	55.90%	55.74%	48.52%	56.68%	68.59%	**89.50%**	+20.91%
	Sum$\mathcal{R}$	↑	38.02	50.48	39.6	58.51	108.46	125.88	128.97	107.63	130.92	163.07	**202.1**	**+39.03**
ITR	$\mathcal{R}$@1	↑	4.010%	4.590%	4.590%	4.290%	20.97%	22.76%	23.96%	19.26%	23.39%	27.99%	**33.10%**	+5.11%
	$\mathcal{R}$@5	↑	11.03%	14.99%	16.50%	14.90%	40.49%	41.89%	46.31%	39.90%	44.67%	60.09%	**77.20%**	+17.11%
	$\mathcal{R}$@10	↑	22.14%	24.10%	26.60%	24.20%	48.21%	50.77%	52.12%	46.05%	52.55%	68.37%	**91.10%**	+22.73%
	Sum$\mathcal{R}$	↑	37.18	43.68	47.69	43.39	109.67	115.42	122.39	105.21	120.61	156.45	**201.4**	**+44.95**
M-CR	$\mathcal{A}$	↑	N/A	N/A	N/A	N/A	N/A	90.77%	91.25%	N/A	91.79%	95.07%	**98.26%**	+3.19%
	macro-$\mathcal{F}$	↑	N/A	N/A	N/A	N/A	N/A	0.699	0.705	N/A	0.727	0.714	**0.896**	**+0.169**
	Sum$\mathcal{C}$	↑	N/A	N/A	N/A	N/A	N/A	160.67	161.75	N/A	164.49	166.47	**187.86**	**+21.39**
S-CR	$\mathcal{A}$	↑	N/A	N/A	N/A	N/A	N/A	80.11%	85.27%	N/A	84.23%	88.07%	**93.57%**	+5.50%
	macro-$\mathcal{F}$	↑	N/A	N/A	N/A	N/A	N/A	0.575	0.620	N/A	0.591	0.636	**0.829**	**+0.193**
	Sum$\mathcal{C}$	↑	N/A	N/A	N/A	N/A	N/A	137.61	147.27	N/A	143.33	151.67	**176.47**	**+24.80**


Short sleeve cotton jersey t-shirt in black. Rib knit crewneck collar. Knot at cuff. Tonal stitching.

Cropped poplin t-shirt in white. Ribbed crewneck collar. Boxy fit. Tonal stitching.

Long sleeve mohair-silk knit sweater in cream. Ribbed crewneck collar, cuffs, and hem. Tonal stitching.

Straight-leg old denim jeans in indigo. Mid-rise. Four-pocket styling. Button-fly. Tonal stitching. Approx. 7.5" leg opening.

- Long sleeve 'fully fashioned French terry' knit cashmere pullover in light grey. Rib knit crewneck collar, cuffs, and hem. Raglan sleeves. Rib knit panel at armscyes and side-seams. Signature 'four bar' striping knit in white at upper sleeve. Signature tricolor grosgrain pull-tab at back yoke. Tonal stitching. (99.96%, Matched)
- Long sleeve rib knit silk and cotton-blend sweater in navy. Crewneck collar. Red trim at hem. Dropped shoulders. Tonal stitching. (99.91%)
- Cotton knit pullover in black. Ribbed crewneck collar, cuffs, and hem. Intarsia skull pattern knit at front in black and grey. Tonal stitching. (85.90%)
- Long sleeve cotton sweater in navy. Rib knit crewneck collar, cuffs, and hem. Raglan sleeves. Tonal stitching. (0.12%)
- Long sleeve rib knit alpaca and wool-blend sweater in off-white. Distressing throughout. Crewneck collar. Tonal stitching. (0.01%)
- Long sleeve 'chunky' knit wool sweater in off-white. Rib knit V-neck collar, cuffs, and hem. Patch pocket at body. Tonal stitching. (99.95%, Matched)
- Long sleeve rib knit cotton off-the-shoulder sweater in black. V-neck collar. Tonal stitching. (99.94%)
- Long sleeve rib knit merino wool off-the-shoulder pullover in black. Off-white rib knit cotton tank top-style underlay at V-neck collar. Tonal stitching. (64.74%)
- Long sleeve knit crepe sweater in black. V-neck collar. Drop-tail hem. Tonal stitching. (0.50%)
- Long sleeve boxy ribbed knit sweater in black. V-neck collar. Trim in white at collar and cuffs. Raglan sleeves. Tonal stitching. (0.03%)

Fig. 5 Visualization results on the TIR and ITR tasks in terms of top-five ranked probabilities predicted by our MVLT. "Matched" indicates the ground-truth image-text pair.

images on the validation part of Fashion-Gen 2018 (a) and our e-commerce website (b). As seen, the reconstruction performance is truly remarkable. Since it requires our method to learn the fashion semantics truly, such results demonstrate the generative ability of our approach.

4.3 Ablation studies

Mask Ratio. Table 3 (a) presents four variants for different mask probability r_v (i.e., 0.10 (A1), 0.30 (A2), 0.70 (A3) and 0.90 (A4)) and our choice: 0.50 (Final). The $\mathcal{R}$@5 rises steadily with the masking probability until it reaches the sweet spot (75.70% → 78.00%); then it reaches performance plummets (73.80%). We argue that increasing the r_v will make MIR more complex, allowing MVLT to learn better semantics in a more restricted situation. However, masking out too much region will naturally result in losing valid visual information, leading to bad results.

Masked unit size. Thanks to PVT's flexibility, we can easily try different sizes of masked patches. As shown in Table 3 (b), we derive four variants with masked unit

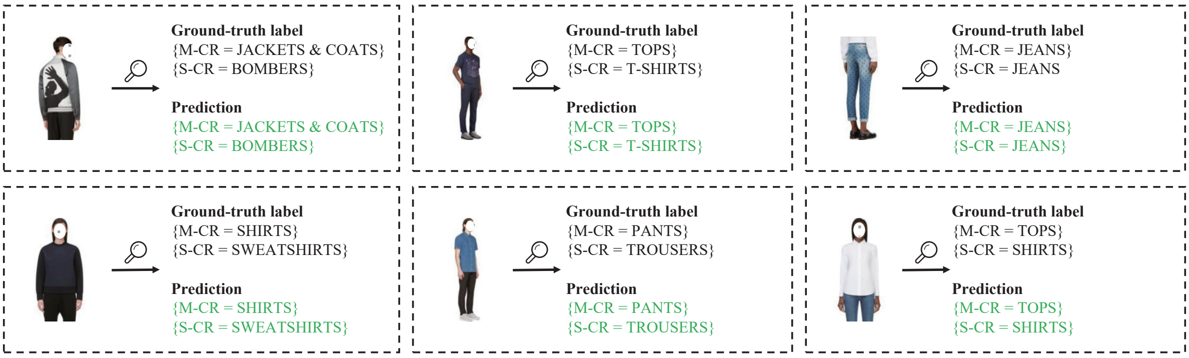
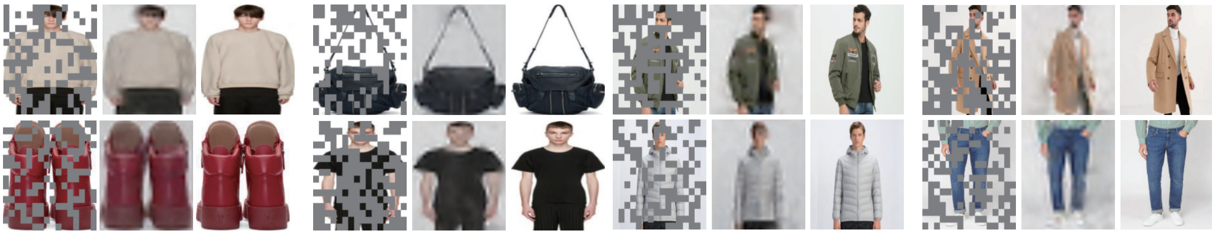


Fig. 6 The visualization of main-/sub-category recognition results on Fashion-Gen. The green predictions hit the targets.


(a) Fashion-gen (In-domain)

(b) Data from e-commercial website (out-of-domain)

Fig. 7 Visualization of samples generated by our MVLT. The gray blocks represent the masked regions.

size α (i.e., 1 (B1), 2 (B2), 8 (B3), 16 (B4)) to compare with our setting: 4 (Final). We found that the performance is sensitive to this factor. It makes sense, revealing how vital it is to learn a robust fashion-related representation with moderate granularity.

Masking style. As shown in Fig. 8, we designed four types of masking strategies for the MIR task, whose quantitative differences are presented in Table 3 (c), i.e., grid (C1), stroke (C2), center (C3) and our random grid (Final) masking strategies. As can be seen, the random grid masking (Final) yields the best results, while the other three perform poorly. We believe this is because, in comparison to the grid (C1) and center (C3), random grid masking (Final) can help MVLT construct comprehensive representations. As our strategy (Final) does, the stroke (C2) also randomly masks the image given, yet it more or less leaves unmasked visual cues in the sub-patches. Our strategy enables the model to easily predict the masked region because semantics in the image are well preserved, enhancing the model's robustness to learning in-sight knowledge.

Pre-training objectives. As shown in Table 3 (d), we derive four different variants from investigating the contribution of each objective, including ITM (D1), ITM+MIR (D2), ITM+MLM (D3), and our ITM+MIR+MLM (Final). When comparing D3 to D1 and D2 in the TIR task, we can see that D3 has a better performance in the $\mathcal{R}$@5 metric: 74.10% (D1) < 76.00% (D2) < 76.20% (D3). We conclude MLM task can help the model thoroughly learn the language knowledge, so it provides a more precise query to recall better-matching images. In the ITR task, we find a similar conclusion when comparing (D2) to (D1) and D3 in $\mathcal{R}$@5 metric: 70.80% (D1) < 75.50% (D2) < 76.30% (D3). It indicates that better visual learning leads to an accurate image query to match the most appropriate caption.

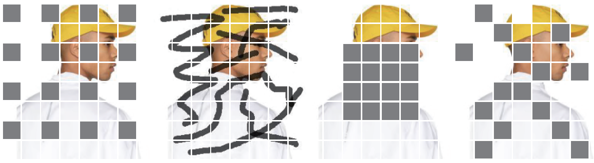
Loading pre-trained weight. As seen in Table 4, we add an experiment to demonstrate it is very important to load the PVT's weight pre-trained on ImageNet[51]. If not, it is obvious that our MVLT will suffer fierce drops (i.e., ITR: 77.20% → 71.50% in $\mathcal{R}$@5, S-CR: 93.57% → 92.90% in $\mathcal{A}$). It is reasonable because a method pre-trained on large-scale general datasets can be more applicable in a specific field. It has already learned information such as color, texture, shape, etc.

4.4 More discussions

How does MVLT perform in general domains? We discuss two extended questions to investigate the potential abilities in general settings further. *1) Can the general models be directly transferred to the fashion domain?* Inspired by the huge impact of general vision-language models, as in Table 5, we further investigate the zero-shot performance of two typical general models (i.e., ViLBERT[14] and CLIP[52]). This has once again demonstrated the necessity and superiority of MVLT pre-trained on specific domains. *2) Can MVLT also work well in the general domain?* We further verify the potential ability of our MVLT in the general domain. Table 6 reports the performance on the MS-COCO 2014 dataset[53], where MVLT follows the same training standards as in [37]. It shows that MVLT achieves promising results compared to the latest models (i.e., Unicoder-VL[54], UNITER[15], and ViLT[37]) without extra training data and special retrieval losses during the training. It indicates that MVLT

Table 3 Ablation studies of five key pre-training factors on our MVLT. More relevant analyses refer to Section 4.3.

		(a) Mask ratio (r_v)				**(b) Masking unit size (α)**				**(c) Masking style**			**(d) Pre-training tasks**			**(e) Pre-train**	**MVLT**
		(A1)	(A2)	(A3)	(A4)	(B1)	(B2)	(B3)	(B4)	(C1)	(C2)	(C3)	(D1)	(D2)	(D3)	(E1)	**(Final)**
App.	Metric	0.10	0.30	0.70	0.90	1	2	8	16	Grid	Stroke	Center	ITM	ITM+MIR	ITM+MLM	*w/o* PVT	
TIR	$\mathcal{R}$@1	31.10%	33.50%	30.50%	30.70%	31.90%	30.30%	30.00%	32.20%	32.20%	31.40%	30.40%	30.40%	32.20%	32.90%	29.00%	**34.60%**
	$\mathcal{R}$@5	75.70%	76.00%	75.50%	73.80%	75.30%	75.60%	73.90%	76.90%	75.30%	76.10%	75.10%	74.10%	76.00%	76.20%	72.20%	**78.00%**
	$\mathcal{R}$@10	88.60%	88.70%	88.00%	88.60%	89.60%	88.60%	88.20%	88.60%	88.50%	89.20%	87.20%	83.50%	87.20%	88.60%	86.60%	**89.50%**
	Sum$\mathcal{R}$	195.40	198.20	194.00	193.10	196.80	194.50	192.10	197.70	196.00	196.70	192.70	188.00	195.40	197.70	187.80	**202.10**
	Diff	−6.70	−3.90	−8.10	−9.00	−5.30	−7.60	−10.00	−4.40	−6.10	−5.40	−9.40	−14.10	−6.70	−4.40	−14.30	–
ITR	$\mathcal{R}$@1	30.00%	29.90%	29.90%	28.50%	29.00%	29.70%	29.00%	28.90%	31.40%	31.10%	30.10%	29.30%	30.40%	28.40%	25.60%	**33.10%**
	$\mathcal{R}$@5	75.70%	74.90%	76.50%	75.00%	76.90%	77.10%	74.20%	77.30%	77.40%	74.50%	73.90%	70.80%	75.50%	76.30%	71.50%	**77.20%**
	$\mathcal{R}$@10	88.80%	89.00%	89.20%	88.20%	89.40%	87.70%	88.00%	89.90%	89.60%	88.50%	87.80%	86.80%	87.80%	88.80%	85.90%	**91.10%**
	Sum$\mathcal{R}$	194.50	193.80	195.60	191.70	195.30	194.50	191.20	196.10	198.40	194.10	191.80	186.90	193.70	193.50	183.00	**201.40**
	Diff	−6.90	−7.60	−5.80	−9.70	−6.10	−6.90	−10.20	−5.30	−3.00	−7.30	−9.60	−14.50	−7.70	−7.90	−18.40	–
M-CR	$\mathcal{A}$	98.16%	97.87%	98.09%	98.06%	98.03%	98.04%	98.11%	98.01%	98.12%	98.07%	98.04%	96.49%	97.11%	98.08%	97.92%	**98.26%**
	macro-$\mathcal{F}$	0.870	0.860	0.890	0.870	0.870	0.880	0.850	0.870	0.869	0.877	0.870	0.806	0.853	0.876	0.879	**0.896**
	Sum$\mathcal{C}$	185.16	183.87	187.09	185.06	185.03	186.04	183.11	185.01	185.02	185.77	185.04	177.09	182.41	185.68	185.82	**187.86**
	Diff	−2.70	−3.99	−0.77	−2.80	−2.83	−1.82	−4.75	−2.85	−2.84	−2.09	−2.82	−10.77	−5.45	−2.18	−2.04	–
S-CR	$\mathcal{A}$	93.10%	93.34%	93.36%	93.23%	93.29%	93.34%	93.32%	93.32%	93.37%	93.21%	93.59%	89.64%	90.87%	93.29%	92.90%	**93.57%**
	macro-$\mathcal{F}$	0.800	0.810	0.820	0.810	0.810	0.810	0.800	0.799	0.794	0.814	0.830	0.703	0.728	0.809	0.790	**0.829**
	Sum$\mathcal{C}$	173.10	174.34	175.36	174.23	174.29	174.34	173.32	173.22	172.77	174.61	176.59	159.94	163.67	174.19	171.90	**176.47**
	Diff	−3.37	−2.13	−1.11	−2.24	−2.18	−2.13	−3.15	−3.25	−3.70	−1.86	+0.12	−16.53	−12.80	−2.28	−4.57	–



(C1) Grid (C2) Stroke (C3) Center Random grid

Fig. 8 We designed four strategies to mask fashion images. The random grid performs the best.

is also a promising solution when extended to general scenes.

Why do pyramid architecture and MIR benefit? As mentioned in the introduction, there are two understudied problems in the fashion domain. To solve the transferability problem, pyramidal architecture[21] takes raw data as input without complex pre-processing, which essentially alleviates the applied burden in industry. Besides, MIR does not need human annotations like classification tags, bounding boxes, or pixel-wise segmentation labels. For the granularity problem[55], the pyramidal architecture[21] provides multi-scale features with rich semantics. Combined with the MIR task, our framework can represent multi-grained fashion knowledge (e.g., dress, V-neck). These features are helpful and urgently required in this field.

A VL model that performs well for semantic understanding tasks (e.g., retrieval[56], classification) can serve as a good foundation and be easily applied to downstream tasks (e.g., text-to-image synthesis[57], image captioning) by utilizing an additional decoder. We did not conduct image captioning experiments because we focused on basic representation learning in fashion this time.

MVLT VS. MAE[34]. MAE learns general representations by allowing the model to explore pixel-to-pixel associations. Therefore, MVLT and MAE are similar in this regard. However, our MVLT is the first that introduces the vision reconstruction-alike pre-training for multi-modal research (e.g., fashion domain).

5 Conclusions

We present a vision-language framework named MVLT, which provides two contributions in this field: 1) a newly-designed masked image reconstruction (MIR) objective and 2) an end-to-end pre-training scheme. The experimental and ablative analysis demonstrates the superiority of various matching and generative tasks. MVLT outperforms the cutting-edge method Kaleido-BERT with large margins on retrieval and recognition tasks, which would catalyze the fashion domain. The designed out-of-box method working end-to-end could simplify the workflow (e.g., data pre-processing and model training) for the actual engineering value, which improves development

Table 4 Ablation study for the contribution of loading PVT's weights pre-trained on ImageNet[51]

	TIR		ITR		M-CR		S-CR	
	$\mathcal{R}$@5	$\mathcal{R}$@10	$\mathcal{R}$@5	$\mathcal{R}$@10	$\mathcal{A}$	macro-$\mathcal{F}$	$\mathcal{A}$	macro-$\mathcal{F}$
w/o PVT	72.20%	86.60%	71.50%	85.90%	97.92%	0.879	92.90%	0.790
w/ PVT	**78.00%**	**89.50%**	**77.20%**	**91.10%**	**98.26%**	**0.896**	**93.57%**	**0.829**
Diff	+5.80%	+2.90%	+5.70%	+5.20%	+0.34%	+1.7%	+0.67%	+3.9%

Table 5 The comparison of zero-shot retrieval results on the Fashion-Gen dataset

	TIR			ITR		
	$\mathcal{R}$@1↑	$\mathcal{R}$@5↑	$\mathcal{R}$@10↑	$\mathcal{R}$@1↑	$\mathcal{R}$@5↑	$\mathcal{R}$@10↑
ViLBERT (Zero-shot)	7.18%	18.73%	29.84%	8.99%	15.34%	26.14%
CLIP (Zero-shot)	16.30%	40.60%	55.60%	13.60%	43.10%	57.60%
MVLT (OUR)	**34.60%**	**78.00%**	**89.50%**	**33.10%**	**77.20%**	**91.10%**

Table 6 Retrieval results on the MS-COCO 2014 dataset. † means using an extra feature extractor (e.g., Faster RCNN).

	TIR task (5K Test)			ITR task (5K Test)		
	$\mathcal{R}$@1↑	$\mathcal{R}$@5↑	$\mathcal{R}$@10↑	$\mathcal{R}$@1↑	$\mathcal{R}$@5↑	$\mathcal{R}$@10↑
Unicoder-VL†	48.40%	76.70%	85.90%	62.30%	87.10%	92.80%
UNITER-Base†	**50.30%**	78.50%	87.20%	64.40%	87.40%	93.10%
ViLT-Base/32	41.30%	72.00%	82.50%	61.80%	86.20%	92.60%
MVLT (OUR)	49.66%	**79.88%**	**87.50%**	**65.38%**	**90.04%**	**93.60%**

and business efficiency on large-scale e-commerce websites by approximately 50%.

In the future, we will continue to investigate an extremely efficient method in this field using famous technologies such as hashing[58], network pruning, and knowledge distillation to alleviate the storage and computing limitations in real-world e-commerce applications.

Acknowledgements

This work is funded by Toyota Motor Europe via the research project TRACE-Zürich. The authors also would like to thank the anonymous reviewers and editor for their helpful comments on this manuscript. Open access funding provided by Swiss Federal Institute of Technology Zurich.

Conflicts of Interests

The authors declared that they have no conflicts of interest in this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Open Access

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. H. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16 ×16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[2] Z. Liu, Y. T. Lin, Y. Cao, H. Hu, Y. X. Wei, Z. Zhang, S. Lin, B. N. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 9992–10002, 2021. DOI: 10.1109/ICCV48922.2021.00986.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, US, pp. 6000–6010, 2017.

[4] T. X. Sun, X. Y. Liu, X. P. Qiu, X. J. Huang. Paradigm shift in natural language processing. *Machine Intelligence Research*, vol. 19, no. 3, pp. 169–183, 2022. DOI: 10.1007/s11633-022-1331-6.

[5] S. Agarwal, G. Krueger, J. Clark, A. Radford, J. W. Kim, M. Brundage. Evaluating CLIP: Towards characterization of broader capabilities and downstream implications, [Online], Available: https://arxiv.org/abs/2108.02818, August 05, 2021.

[6] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, I. Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, Article number 233, 2020.

[7] J. Y. Lin, R. Men, A. Yang, C. Zhou, Y. C. Zhang, P. Wang, J. R. Zhou, J. Tang, H. X. Yang. M6: Multi-modality-to-multi-modality multitask mega-transformer for unified pretraining. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021. DOI: 10.1145/3447548.3467206.

[8] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8821–8831, 2021.

[9] H. Wu, Y. P. Gao, X. X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, R. Feris. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 11302–11312, 2021. DOI: 10.1109/CVPR46437.2021.01115.

[10] J. Devlin, M. W. Chang, K. Lee, K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, USA, pp. 4171–4186, 2019. DOI: 10.18653/v1/N19-1423.

[11] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[12] S. Q. Ren, K. M. He, R. Girshick, J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of 2015 Annual Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 91–99, 2015.

[13] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, A. Sacheti. ImageBERT: Cross-modal pre-training with large-scale weak-supervised image-text data, [Online], Available: https://arxiv.org/abs/2001.07966, January 23, 2020.

[14] J. S. Lu, D. Batra, D. Parikh, S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 13–23, 2019.

[15] Y. C. Chen, L. J. Li, L. C. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. J. Liu. UNITER: UNiversal image-TExt representation learning. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glas-

gow, UK, pp. 104–120, 2020. DOI: 10.1007/978-3-030-58577-8_7.

[16] W. L. Hsiao, I. Katsman, C. Y. Wu, D. Parikh, K. Grauman. Fashion++: Minimal edits for outfit improvement. In *Proceedings of IEEE/CVF International Conference On Computer Vision*, IEEE, Montreal, Canada, pp. 5046–5055, 2019. DOI: 10.1109/ICCV.2019.00515.

[17] M. I. Vasileva, B. A. Plummer, K. Dusad, S. Rajpal, R. Kumar, D. Forsyth. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 405–421, 2018. DOI: 10.1007/978-3-030-01270-0_24.

[18] D. P. Fan, M. C. Zhuge, L. Shao. Domain Specific Pre-Training of Cross Modality Transformer Model, US20220277218, September 2022.

[19] D. H. Gao, L. B. Jin, B. Chen, M. H. Qiu, P. Li, Y. Wei, Y. Hu, H. 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*, ACM, pp. 2251–2260, 2020. DOI: 10.1145/3397271.3401430.

[20] M. C. Zhuge, D. H. Gao, D. P. Fan, L. B. Jin, B. Chen, H. M. Zhou, M. H. Qiu, L. Shao. Kaleido-BERT: Vision-language pre-training on fashion domain. In *Proceedings of IEEE/CVF Conference on computer vision and pattern recognition*, IEEE, Nashville, USA, pp. 12642–12652, 2021. DOI: 10.1109/CVPR46437.2021.01246.

[21] W. H. Wang, E. Z. Xie, X. Li, D. P. Fan, K. T. Song, D. Liang, T. Lu, P. Luo, L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 548–558, 2021. DOI: 10.1109/ICCV48922.2021.00061.

[22] X. W. Yang, H. M. Zhang, D. Jin, Y. R. Liu, C. H. Wu, J. C. Tan, D. L. Xie, J. Wang, X. Wang. Fashion captioning: Towards generating accurate descriptions with semantic rewards. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 1–17, 2020. DOI: 10.1007/978-3-030-58601-0_1.

[23] Z. Al-Halah, K. Grauman. From Paris to Berlin: Discovering fashion style influences around the world. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 10133–10142, 2020. DOI: 10.1109/CVPR42600.2020.01015.

[24] H. Tan, M. Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, pp. 5100–5111, 2019. DOI: 10.18653/v1/D19-1514.

[25] W. J. Su, X. Z. Zhu, Y. Cao, B. Li, L. W. Lu, F. R. Wei, J. F. Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *Proceedings of the 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia, 2020.

[26] K. H. Lee, X. Chen, G. Hua, H. D. Hu, X. D. He. Stacked cross attention for image-text matching. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 212–228, 2018. DOI: 10.1007/978-3-030-01225-0_13.

[27] Z. X. Niu, M. Zhou, L. Wang, X. B. Gao, G. Hua. Hierarchical multimodal LSTM for dense visual-semantic embedding. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 1899–1907, 2017. DOI: 10.1109/ICCV.2017.208.

[28] J. Xia, M. Zhuge, T. Geng, S. Fan, Y. Wei, Z. He, F. Zheng. Skating-mixer: Multimodal MLP for scoring figure skating, [Online], Available: https://arxiv.org/abs/2203.03990, 2022.

[29] X. J. Li, X. Yin, C. Y. Li, P. C. Zhang, X. W. Hu, L. Zhang, L. J. Wang, H. D. Hu, L. Dong, F. R. Wei, Y. J. Choi, J. F. Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 121–137, 2020. DOI: 10.1007/978-3-030-58577-8_8.

[30] M. C. Zhuge, D. P. Fan, N. Liu, D. W. Zhang, D. Xu, L. Shao. Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to be published. DOI: 10.1109/TPAMI.2022.3179526.

[31] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, pp. 2048–2057, 2015.

[32] T. Arici, M. S. Seyfioglu, T. Neiman, Y. Xu, S. Train, T. Chilimbi, B. Zeng, I. Tutar. MLIM: Vision-and-language model pre-training with masked language and image modeling, [Online], Available: https://arxiv.org/abs/2109.12178, September 24, 2021.

[33] H. B. Bao, L. Dong, S. L. Piao, F. R. Wei. BEiT: BERT pre-training of image transformers. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

[34] K. M. He, X. L. Chen, S. N. Xie, Y. H. Li, P. Dollár, R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, USA, pp. 15979–15988, 2022. DOI: 10.1109/CVPR52688.2022.01553.

[35] Z. C. Huang, Z. Y. Zeng, B. Liu, D. M. Fu, J. L. Fu. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers, [Online], Available: https://arxiv.org/abs/2004.00849, June 22, 2020.

[36] X. D. Lin, G. Bertasius, J. Wang, S. F. Chang, D. Parikh, L. Torresani. VX2TEXT: End-to-end learning of video-based text generation from multimodal inputs. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 7001–7011, 2021. DOI: 10.1109/CVPR46437.2021.00693.

[37] W. Kim, B. Son, I. Kim. ViLT: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5583–5594, 2021.

[38] M. Yan, H. Y. Xu, C. L. Li, B. Bi, J. F. Tian, M. Gui, W. Wang. Grid-VLP: Revisiting grid features for vision-language pre-training, [Online], Available: https://arxiv.org/abs/2108.09479, August 21, 2021.

[39] Z. C. Huang, Z. Y. Zeng, Y. P. Huang, B. Liu, D. M. Fu, J. L. Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Nashville, USA, pp. 12971–12980, 2021. DOI: 10.1109/CVPR46437.2021.01278.

[40] S. Goenka, Z. H. Zheng, A. Jaiswal, R. Chada, Y. Wu, V. Hedau, P. Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings on Conference on computer vision and pattern recognition*, IEEE, New Orleans, USA, pp. 14085–14095, 2022. DOI: 10.1109/CVPR52688.2022.01371.

[41] J. Lei, L. J. Li, L. W. Zhou, Z. Gan, T. L. Berg, M. Bansal, J. J. Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 7327–7337, 2021. DOI: 10.1109/CVPR46437.2021.00725.

[42] H. Y. Xu, M. Yan, C. L. Li, B. Bi, S. F. Huang, W. M. Xiao, F. Huang. E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 503–513, 2021.

[43] H. Akbari, L. Z. Yuan, R. Qian, W. H. Chuang, S. F. Chang, Y. Cui, B. Q. Gong. VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, pp. 24206–24221, 2021.

[44] X. Y. Yi, J. Yang, L. C. Hong, D. Z. Cheng, L. Heldt, A. Kumthekar, Z. Zhao, L. Wei, E. Chi. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, ACM, Copenhagen, Denmark, pp. 269–277, 2019. DOI: 10.1145/3298689.3346996.

[45] O. Ronneberger, P. Fischer, T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Munich, Germany, pp. 234–241, 2015. DOI: 10.1007/978-3-319-24574-4_28.

[46] C. Alberti, J. Ling, M. Collins, D. Reitter. Fusion of detected objects in text for visual question answering. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, pp. 2131–2140, 2019. DOI: 10.18653/v1/D19-1219.

[47] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, C. Pal. Fashion-gen: The generative fashion dataset and challenge, [Online], Available: https://arxiv.org/abs/1806.08317v1, July 30, 2018.

[48] R. Kiros, R. Salakhutdinov, R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models, [Online], Available: https://arxiv.org/abs/1411.2539, 2014.

[49] F. Faghri, D. J. Fleet, J. R. Kiros, S. Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of British Machine Vision Conference*, Newcastle, UK, 2018.

[50] Y. X. Wang, H. Yang, X. M. Qian, L. Ma, J. Lu, B. Li, X. Fan. Position focused attention network for image-text matching. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, Macao, China, pp. 3792–3798, 2019.

[51] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, F. F. Li. ImageNet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Miami, USA, pp. 248–255, 2009. DOI: 10.1109/CVPR.2009.5206848.

[52] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

[53] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the 13th European Conference on Computer Vision*, Springer, Zürich, Switzerland, pp. 740–755, 2014. DOI: 10.1007/978-3-319-10602-1_48.

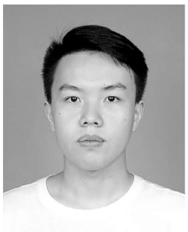
[54] G. Li, N. Duan, Y. J. Fang, M. Gong, D. Jiang. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence, the 32nd Innovative Applications of Artificial Intelligence Conference, the 10th AAAI Symposium on Educational Advances in Artificial Intelligence*, New York, USA, pp. 11336–11344, 2020.

[55] L. Wu, D. Y. Liu, X. J. Guo, R. C. Hong, L. C. Liu, R. Zhang. Multi-scale spatial representation learning via recursive hermite polynomial networks. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, Vienna, Austria, pp. 1465–1473, 2022. DOI: 10.24963/ijcai.2022/204.

[56] D. P. Chen, M. Wang, H. B. Chen, L. Wu, J. Qin, W. Peng. Cross-modal retrieval with heterogeneous graph embedding. In *Proceedings of the 30th ACM International Conference on Multimedia*, ACM, Lisboa, Portugal, pp. 3291–3300, 2022. DOI: 10.1145/3503161.3548195.

[57] D. Y. Liu, L. Wu, F. Zheng, L. Q. Liu, M. Wang. Verbal-person nets: Pose-guided multi-granularity language-to-person generation. *IEEE Transactions on Neural Networks and Learning Systems*, to be published. DOI: 10.1109/TNNLS.2022.3151631.

[58] Z. Zhang, H. Y. Luo, L. Zhu, G. M. Lu, H. T. Shen. Modality-invariant asymmetric networks for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, to be published. DOI: 10.1109/TKDE.2022.3144352.



Ge-Peng Ji received the M. Sc. degree in communication and information systems from Wuhan University, China in 2021. He is currently a Ph. D. degree candidate at Australian National University, supervised by Professor Nick Barnes, majoring in engineering and computer science. He has published about 10 peer-reviewed journal and conference papers. In 2021, he received the Student Travel Award from Medical Image Computing and Computer-assisted Intervention Society.

His research interests lie in computer vision, especially in a variety of dense prediction tasks, such as video analysis, medical image segmentation, camouflaged object segmentation, and saliency detection.

E-mail: gepengai.ji@gmail.com

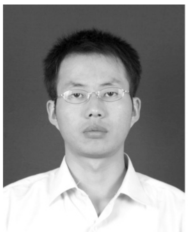
ORCID iD: 0000-0001-7092-2877

Mingchen Zhuge received the M. Sc. degree in computer science from China University of Geosciences, China in 2021. He is a Ph. D. degree candidate in King Abdullah University of Science and Technology (KAUST) under the supervision of Prof. Juergen Schmidhuber. In 2019, he won the championship in the ZTE algorithm competition. He has worked as an intern at Alibaba Group and IIAI, as well as a visiting scholar at SUSTech. Besides, he has been invited to serve as a top conference reviewer for CVPR, ICML, ECCV, NeurIPS, etc.

His research interests include multi-modal learning and reinforcement learning.

E-mail: mczhuge@gmail.com

ORCID iD: 0000-0003-2561-7712

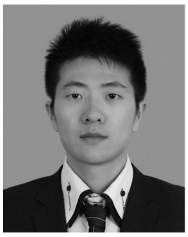


Dehong Gao received the Ph. D. degree from The Hong Kong Polytechnic University, China in 2014. He is now working as an associate professor in Northwestern Polytechnical University, China.

His research interests include information retrieval, recommendation, natural language processing and machine learning.

E-mail: gaodehong_polyu@163.com, dehong.gdh@alibaba-inc.com

ORCID iD: 0000-0002-6636-5702

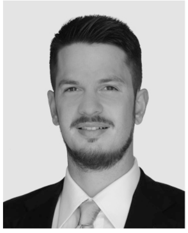


Deng-Ping Fan received the Ph. D. degree from the Nankai University, China in 2019. He joined Inception Institute of Artificial Intelligence (IIAI), UAE in 2019. He has published about 50 top journal and conference papers such as TPAMI, IJCV, TIP, TNNLS, TMI, CVPR, ICCV, ECCV, IJCAI, etc. He won the Best Paper Finalist Award at IEEE CVPR 2019, the Best Paper Award Nominee at IEEE CVPR 2020. He was recognized as the CVPR 2019 outstanding reviewer with a special mention award, the CVPR 2020 outstanding reviewer, the ECCV 2020 high-quality reviewer, and the CVPR 2021 outstanding reviewer. He served as a program committee board (PCB) member of IJCAI 2022−2024, a senior program committee (SPC) member of IJCAI 2021, a program committee members (PC) of CAD&CG 2021, a committee member of China Society of Image and Graphics (CSIG), area chair in NeurIPS 2021 Datasets and Benchmarks Track, area chair in MICCAI2020 Workshop.

His research interests include computer vision, deep learning, and visual attention, especially the human vision on co-salient object detection, RGB salient object detection, RGB-D salient object detection, and video salient object detection.

E-mail: dengpfan@gmail.com (Corresponding author)

ORCID iD: 0000-0002-5245-7518

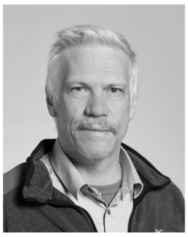


Christos Sakaridis received the M. Sc. degree in computer science from ETH Zürich, Switzerland in 2016 and his Diploma in electrical and computer engineering from the National Technical University of Athens, Greece in 2014, conducting his Diploma thesis at CVSP Group under the supervision of Prof. Petros Maragos. He received the Ph.D. degree in electrical engineering and information technology from ETH Zürich, Switzerland in 2021, working at Computer Vision Lab and supervised by Prof. Luc Van Gool. He is a postdoctoral researcher at Computer Vision Lab, ETH Zürich, Switzerland. Since 2021, he is the Principal Engineer in TRACE-Zürich, a project on computer vision for autonomous cars running at Computer Vision Lab and funded by Toyota Motor Europe. Moreover, he is the team leader in the EFCL project Sensor Fusion, in which they develop adaptive sensor fusion architectures for high-level visual perception.

His broad research fields are computer vision and machine learning. The focus of his research is on high-level visual perception, involving adverse visual conditions, domain adaptation, semantic segmentation, depth estimation, object detection, synthetic data generation, and fusion of multiple sensors (including lidar, radar and event cameras, with emphasis on their application to autonomous cars and robots).

E-mail: csakarid@vision.ee.ethz.ch

ORCID iD: 0000-0003-1127-8887



Luc Van Gool received the Ph. D. degree in electromechanical engineering at Katholieke Universiteit Leuven, Belgium in 1981. Currently, he is a professor at Katholieke Universiteit Leuven, in Belgium and ETH Zürich, Switzerland. He leads computer vision research at both places and also teaches at both. He has been a program committee member of several major computer vision conferences. He received several Best Paper awards, won a David Marr Prize and a Koenderink Award, and was nominated Distinguished Researcher by the IEEE Computer Science committee. He is a co-founder of 10 spin-off companies.

His interests include 3D reconstruction and modeling, object recognition, tracking, gesture analysis, and a combination of those.

E-mail: vangool@vision.ee.ethz.ch

ORCID iD: 0000-0002-3445-5711