

Automatic Generation of Product-Image Sequence in E-commerce

Xiaochuan Fan

JD.COM Research

Mountain View, USA

efan3000@gmail.com

Yue Shang

JD.COM Research

Mountain View, USA

shangyue1230@gmail.com

Yun Xiao

JD.COM Research

Mountain View, USA

yun.xiao@yahoo.com

Chi Zhang

JD.COM Research

Mountain View, USA

zhangc11msu@gmail.com

Xueying Zhang

JD.COM Research

Mountain View, USA

zxycontact@gmail.com

Bo Long

JD.COM

Beijing, China

bo.long@jd.com

Yong Yang

JD.COM

Beijing, China

yangyong280@jd.com

Zhen He

JD.COM

Beijing, China

bjhezhen@jd.com

Lingfei Wu

JD.COM Research

Mountain View, USA

lwu@email.wm.edu

ABSTRACT

Product images are essential for providing desirable user experience in an e-commerce platform. For a platform with billions of products, it is extremely time-costly and labor-expensive to manually pick and organize qualified images. Furthermore, there are the numerous and complicated image rules that a product image needs to comply in order to be generated/selected. To address these challenges, in this paper, we present a new learning framework in order to achieve Automatic Generation of Product-Image Sequence (AGPIS) in e-commerce. To this end, we propose a Multi-modality Unified Image-sequence Classifier (MUIsC), which is able to simultaneously detect all categories of rule violations through learning. MUIsC leverages textual review feedback as the additional training target and utilizes product textual description to provide extra semantic information. Based on offline evaluations, we show that the proposed MUIsC significantly outperforms various baselines. Besides MUIsC, we also integrate some other important modules in the proposed framework, such as primary image selection, non-compliant content detection, and image deduplication. With all these modules, our framework works effectively and efficiently in JD.com recommendation platform. By Dec 2021, our AGPIS framework has generated high-standard images for about 1.5 million products and achieves 13.6% in reject rate. Code of this work is available at <https://github.com/efan3000/muisc>.

CCS CONCEPTS

- Applied computing → Online shopping; • Information systems → Display advertising.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '22, August 14–18, 2022, Washington, DC, USA

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-9385-0/22/08...\$15.00

<https://doi.org/10.1145/3534678.3539149>

KEYWORDS

product image selection, image-sequence classifier, multi-modality fusion, e-commerce

ACM Reference Format:

Xiaochuan Fan, Chi Zhang, Yong Yang, Yue Shang, Xueying Zhang, Zhen He, Yun Xiao, Bo Long, and Lingfei Wu. 2022. Automatic Generation of Product-Image Sequence in E-commerce. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3534678.3539149>

1 INTRODUCTION

Product images are essential for the success of online shopping [3][7][38], as good product images enhance product description and help fill the gap between offline shopping and online shopping. Most of well-known e-commerce platforms have their own various complicated standards on product images, such as Amazon, Ebay, Alibaba and JD.com. "Haohuo" channel¹ (referring to "Discovery Goods"), an important traffic entrance on the top of JD.com (both website and App), is featured by its high standard of product presenting. A typical process for selecting product images are following this procedure: i) "HaoHuo" channel motivates human professions to submit high-quality product images; ii) human reviewers then review the submitted images according to a set of complicated image standards, which are largely designed based on in-house experts' experience and A/B tests. Figure 1 exhibits an example of comparison between product images from JD.com's main site and "Haohuo" channel. We can see that "Haohuo" channel presents the product in a concise and well-organized way. However, for an e-commerce platform with large-scaled product catalogs, it is extremely time-costly and labor-expensive to manually pick product images. In this paper, we propose a novel learning framework for Automatic Generation of Product-Image Sequence (AGPIS), which automatically picks a sequence of product images from candidate images according to a set of numerous and complicated rules.

¹<https://fxhh.jd.com/>

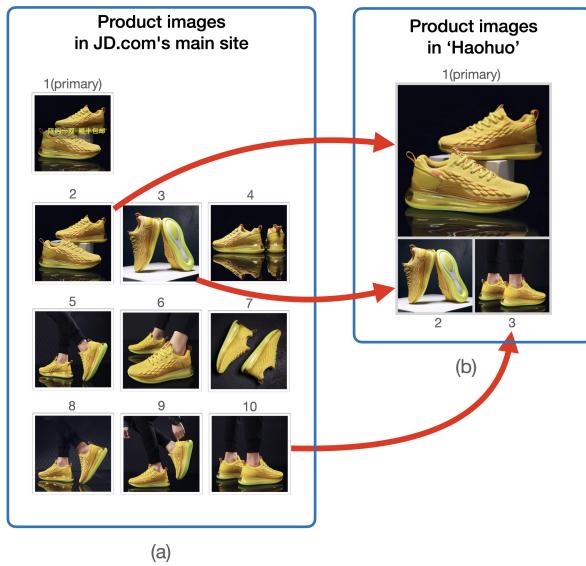


Figure 1: A pair of shoes are presented by images from (a) JD.com's main site and (b) JD.com's "Haohuo" Channel. (a) contains redundant and non-compliant content, while (b) presents all aspects of the product via three images.

There are some early works to address similar problems [2][9], which however are inadequate to address the aforementioned problems due to various unique challenges. Firstly, there are numerous and complicated rules and these rules could change over the time. This makes it unaffordable to develop rule-based specific methods or datasets for each rule due to the huge computation as well as the development and maintenance costs. Furthermore, the existing methods only focus on one or several rules [2][9]. As a result, it is hard to work well with these systems since they are designed and customized with the fixed prior knowledge of rules. For example, detection of non-compliant content often involves logo detection [14][29] and skin region detection [36][15].

Secondly, different rules may require different information. Product image rules can be generally divided into three categories as follows. First, **Single-image rules**. Typical single-image rules are about image quality, e.g. unnatural artifact or blurry image, and non-compliant content, e.g. logo, banner, and water-print. The detection of single-image rule violations is only relevant to an individual image. Second, **Image-pair rules**. Image-pair rules involve matching and comparison of two images in order to avoid redundant or wrong information. An example of image-pair rules states that two images should not present product appearance from similar viewing angles. The detection of this category of violations requires information from a pair of images. Third, **Multi-image rules**. These rules are usually designed for the layout of product images to make sure product information is adequate and released in a proper order. The detection of multi-image rule violations may require information from multiple images or even cross-modality product description. Most of the existing methods [9][14][29][36][15] only focus on violation detection of single-image and image-pair rules, and ignore the exploration of relations between all the images in a sequence.

However, such relations are essential for automatic image selections since a product is presented by multiple images as a whole.

Last but not least, the rich information of image review feedback can be used for automatic image selection. The existing methods usually cast their problems into image classification problems. Then these product images are labeled as either qualified or not qualified according to the rules. However, textual feedback from human reviewers may contain rich semantic information which cannot be easily converted to classification labels. Taking "Haohuo" channel as an example, besides the violated rule name, a review feedback may also include extra information to explain these rejections, including which image in a sequence violates the rule, the location of non-compliant content in an image or what the non-compliant content is. Therefore, such rich semantic information is also helpful for improving automatic image selection.

To address these aforementioned challenges, in this paper, we present a novel learning framework in order to achieve Automatic Generation of Product-Image Sequence (AGPIS) in e-commerce. The core module of our framework is a Multi-modality Unified Image-sequence Classifier (MUIsC), which is able to simultaneously detect all categories of rule violations through learning. Firstly, to obtain adequate information for AGPIS, MUIsC takes as input an image sequence, rather than a single image or a pair of images, and extracts features via a hierarchical encoder. Then, along with the classification task, we use Natural Language Generation (NLG) of image review feedback as an auxiliary training task to fully exploit the rich semantic information of a review feedback. Different from some traditional tasks such as visual question answering [1] and image captioning [22, 34], which aim to generate text conditioned on a visual input, our NLG task functions as a guide for MUIsC to better understand complicated rules during training. In addition, the introduction of NLG task does not incur any additional burden in data processing since no manual labeling is needed. This is especially important for frequent model update in a real application. Lastly, textual product description is also fed to MUIsC as a input to assist image recognition. This requires our proposed MUIsC to be able to perform image-text interaction effectively. Interestingly, the resulting MUIsC has similar input and output with a human reviewer, which means that no prior knowledge or rule-based task is involved in our model.

In order to accumulate data for MUIsC training and make our framework work efficiently, we also integrate additional modules for single-image and image-pair recognition to detect unqualified images and build a sequence candidate for MUIsC. Figure 2 exhibits the overall pipeline of the proposed framework. Given a set of candidate images and textual product description, our framework outputs an image sequence and its probability of being qualified. If the probability is larger than a threshold, we submit the image sequence to a human reviewer and then "Haohuo" channel if approved. The red arrows in Figure 1 indicates an example of the correspondence between candidate images (from JD's main site) and the resulting image sequence of our framework for "Haohuo" channel.

Since its deployment in JD.com in Feb 2021, our AGPIS framework has generated high-standard images for about 1.5 million products and achieves 13.6% in reject rate.

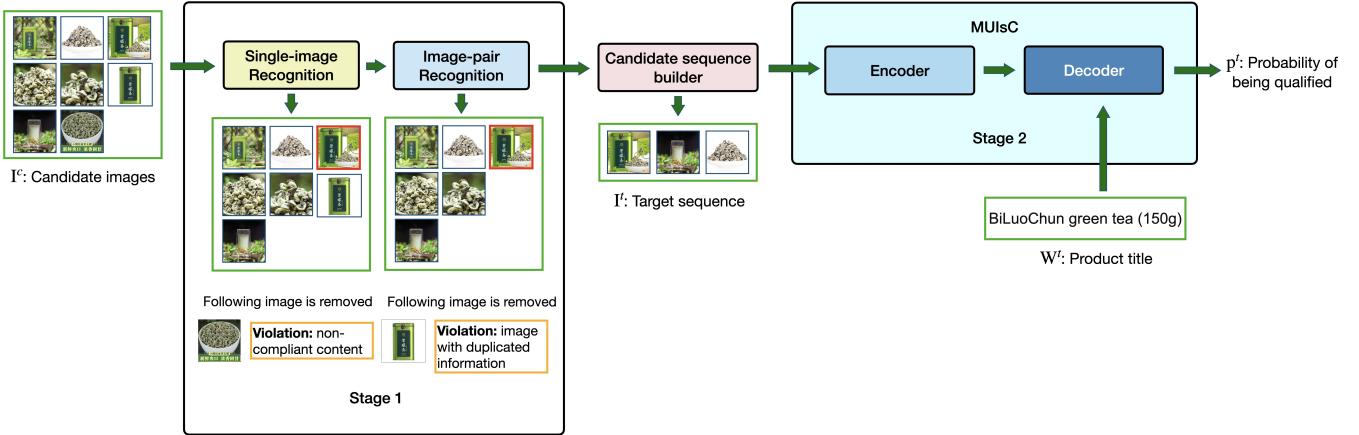


Figure 2: Pipeline of our AGPIS framework. Stage 1 detects non-compliant/redundant images and selects the primary image as shown by the red image border. Given a target image sequence built from stage 1, stage 2 outputs its probability of being qualified.

2 RELATED WORKS

In this section, we introduce works on similar topics and related domains.

E-commerce image selection. The significance of images in e-commerce has been well-studied [3][7][38], but e-commerce retailers that offer marketplace platforms still struggle to control image quality. Gandhi et al. [9] resolves the issue of non-compliant content by combining state-of-the-art image classification and object detection models, but non-compliant content is only part of numerous and complicated rules for automatic image selection. Chaudhuri et al. [2] proposes a system that aggregates images from various suppliers to produce a image set, arranged in an order according to a set of manually-crafted templates. The template designer has to design different templates for different product categories. This costs huge human efforts and leads to low production efficiency. More importantly, this method limits the diversity of visual product presenting, i.e. the products in the same category are presented in a similar style. By contrast, our proposed framework learns to organize images from a large-scaled reviewed data, which is much more flexible. Besides, in [9] and [2], systems consist of a sequence of modules designed for every specific rule, while our single MUIsC model is able to detect all categories of rule violations.

Image aesthetic and quality assessment has received considerable attention in recent years. This technique, which is similar to violation detection of single-image rules in automatic image selection, has been widely used in user album photo selection [19] and image recommendation [37]. For conventional image quality assessment methods, hand-craft features created from either photography practices or objective quality criteria are widely used [5][17][25]. More recently, learnt feature representations using deep neural network has surpassed the performance of hand-craft ones [16][13][30]. Most of the image assessment works are reported on datasets with assessment scores from peer reviewers, and cast the problem into image classification or ranking. Comparatively, textual product description and review feedback are in the form of

natural language and have much richer semantic information than numeric scores. Besides, our framework confronts more complicated image rules and handles image sequences rather than single images.

Vision-and-language (VL) models, leveraging information from both modalities, has been a very active topic recently. Since transformer[33] based models are adopted in computer vision, VL models achieved great success in tasks such as Visual Question Answering (VQA) [1], image captioning [34][22], and image-text matching [20], etc. Most of these works follow the model architecture in variants of visual backbones (ResNet[11], ViT[8], CLIP[26]), text encoder/decoder (Bert[6], Roberta[23], GPT[27]), modality-fusion scheme (single-stream[21], multi-stream[24][12]), and pre-training objectives (masked language modeling[6], masked image modeling [4][31], multimodal alignment[24, 26, 31]). Our MUIsC model also follows the transformer based text-image modeling framework. Specifically we use a transformer-based encoder for visual feature extraction and adopt an encoder-decoder (two-stream) architecture[33] to fuse the visual and text information encoded by separate encoders. Pre-trained visual and language models are also used for model parameter initialization.

3 METHOD

3.1 Data and Notations

In this paper, we propose a framework for Automatic Generation of Product-Image Sequence (AGPIS), which picks a sequence of images from a set of candidate images according to a set of rules. Our work requires data from two different sources, JD.com's main site and JD.com's "HaoHuo" channel. From JD.com's main site, we collect product images as candidates and product title as a textual product description. From "HaoHuo" channel, we collect reviewed image sequences and corresponding textual feedback for model training and evaluation.

We adopt the following notations for our data. A set of K^c candidate images are represented by $I^c = \{i_1^c, \dots, i_{K^c}^c\}$, where $i \in \mathbb{R}^{H \times W \times 3}$

denotes a RGB image, H and W are height and width of an image. Similarly, an image-sequence that we aim to generate, called target image sequence, is represented by $\mathbf{I}^t = \{\mathbf{i}_1^t, \dots, \mathbf{i}_{K^t}^t\}$, where $\mathbf{i}^t \in \mathbf{I}^c$ and $K^t \leq K^c$ is the number of images in \mathbf{I}^t . Taking the dataset used in this work as an example, a product in JD.com's main site has about 7 images on average, i.e. $K^c \approx 7$, and a target sequence consists of 3 images, i.e. $K^t = 3$. Note that our framework can be also applied to the problem with various values of K^t and K^c . Textual feedback and product title can both be represented by a sequence of words, $\mathbf{W}^f = \{\mathbf{w}_1^f, \dots, \mathbf{w}_{K^f}^f\}$ and $\mathbf{W}^t = \{\mathbf{w}_1^t, \dots, \mathbf{w}_{K^t}^t\}$, respectively. For accepted image sequences, we set their textual feedback to a constant word, e.g. "yes", i.e. $K^f = 1$ and $w_1^f = \text{"yes"}$.

3.2 Framework Overview

Figure 2 provides the overview of our proposed framework for automatic image selection. The framework consists of two stages. Stage 1 consists of a single-image recognition module and an image-pair recognition module. Given a set of candidate images \mathbf{I}^c , the single-image recognition module selects the primary (first) image for the target sequence and detects non-compliant content. Following the single-image recognition module, the image-pair recognition module detects the violations of image-pair rules. Then, a target image sequence \mathbf{I}^t is built from the remaining images in \mathbf{I}^c . In stage 2, \mathbf{I}^t and its corresponding textual description \mathbf{W}^t are fed into a model named Multi-modality Unified Image-sequence Classifier (MUIsC), which estimates the probability of \mathbf{I}^t being qualified, denoted as p^t . \mathbf{I}^t and p^t are the final output of our framework. If p^t is larger than a threshold th^t , we send \mathbf{I}^t to a human reviewer and then "Haohuo" channel if approved.

3.3 Stage 1

3.3.1 Single-image Recognition. The single-image recognition module consists of two learning-based models. One selects the primary image for \mathbf{I}^t since some rules are specifically designed for primary image, and the other model detects non-compliant images. Each of these two models works as a binary single-image classifier based on Deep Neural Networks (DNNs).

For the learning of primary-image selection model, we collect data from image sequences approved by human reviewers \mathbf{I}^{t*} and their candidate images \mathbf{I}^{c*} . We label an image, denoted as $\mathbf{i} \in \mathbf{I}^{t*} \cup \mathbf{I}^{c*}$, to $j^{primary}$ as follows,

$$j^{primary} = \begin{cases} 1 & \mathbf{i} = \mathbf{i}_1^{t*} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

, where \mathbf{i}_1^{t*} is the primary image of \mathbf{I}^{t*} .

We train the model on a binary classification task using a cross-entropy loss function. Let the model's output be $p^{primary}$, which is the probability that \mathbf{i} is a primary image. During inference, the image with the largest $p^{primary}$ in \mathbf{I}^c is selected as the primary image, and $p^{primary}$ should be larger than a threshold. If no such image exists, the process of AGPIS is terminated and the framework outputs nothing.

Our model for non-compliant image detection has similar architecture and loss function with the primary-image selection model but uses a different dataset. From reviewed image sequences, we



Figure 3: An example image pair that violates the rule "images with duplicated content". The green boxes show the duplicated patch-pair detected by our method.

picked the ones which are rejected due to non-compliant content. Then the images in these sequences are manually labeled into two groups: compliance and non-compliance. Let the model's output for non-compliant image detection be p^{nc} , which is the probability that an image contains non-compliant content. All images with p^{nc} larger than a threshold are removed from \mathbf{I}^c in the inference stage. If the number of remaining images is less than K^t , the process of AGPIS is terminated and the framework outputs nothing.

3.3.2 Image-pair Recognition. The module of image-pair recognition detects the violations of image-pair rules. In the real applications, this module is required to match and compare a pair of image patches rather than the whole images. Figure 3 shows a pair of images which violate one of our image-pair rules. We can see that these two images contain much duplicated visual information although they do not have exactly the same content. Therefore, a main difficulty in detection of this category of violations is the construction of a set of region proposals which contain meaningful product information. The representative region proposal algorithms include two-stage object detectors (e.g. Faster RCNN [28]), Selective Search [32] and EdgeBoxes [40]. We empirically find that two-stage object detectors perform badly on e-commerce product images, and speculate that this is caused by different data distribution. In our method, we adopt EdgeBoxes to produce a certain number of region proposals for each image in \mathbf{I}^c , and use a pretrained DNN to extract patch features. Then patches from two images are matched and compared based on k nearest neighbor algorithm. In Figure 3, a pair of detected duplicated patches are outlined by green boxes. If a violation is detected in an image pair, the image with smaller p^{nc} or the selected primary image is kept while the other image is removed. If the number of remaining images is less than K^t , our framework outputs nothing.

Based on the output images of stage 1, a target sequence \mathbf{I}^t is built. The selected primary image is the first image of \mathbf{I}^t . Then, we randomly selected $K^t - 1$ images from the remaining images.

3.4 Stage 2: MUIsC

Our Multi-modality Unified Image-sequence Classifier (MUIsC) is designed following the transformer encoder-decoder architecture illustrated in Figure 4. The encoder extracts features for a given image sequence using a hierarchical architecture, while the decoder performs vision-language fusion, and estimates textual feedback and classification probabilities.

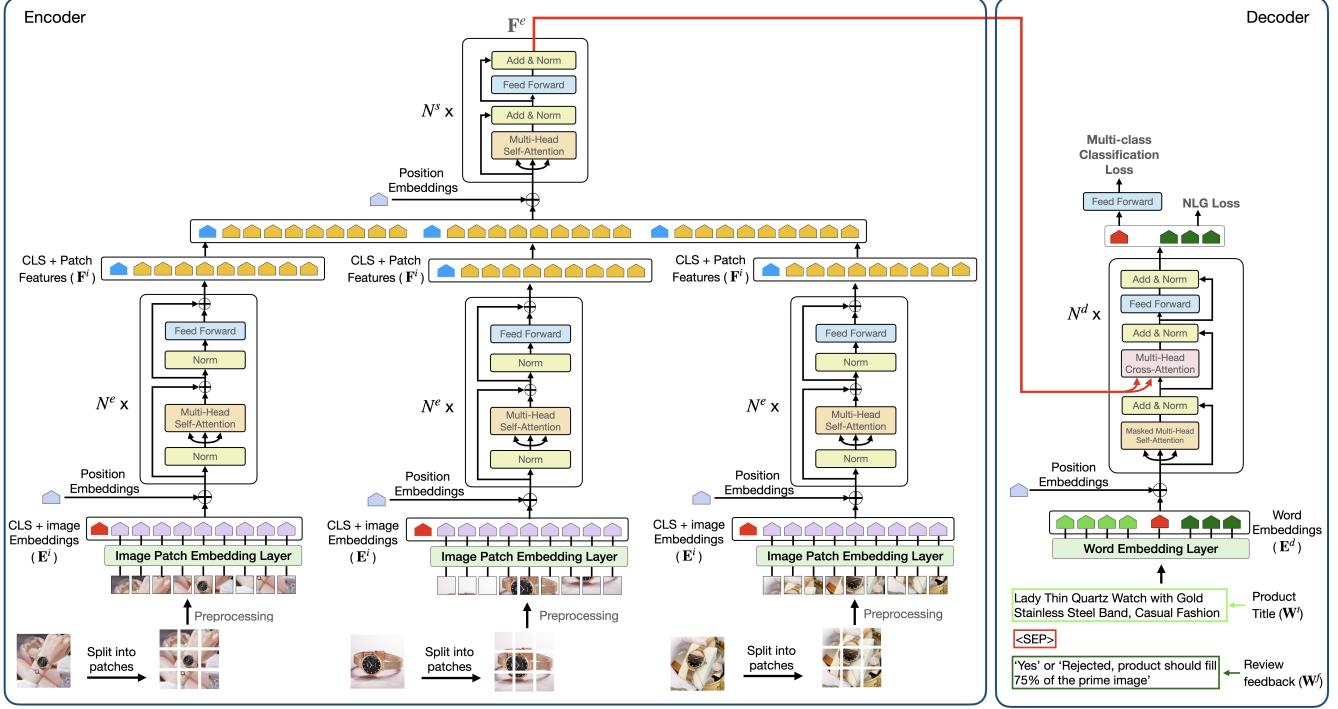


Figure 4: Architecture of MUIsC

3.4.1 Encoder. Given \mathbf{I}^t built after stage 1, the encoder firstly uses a Vision Transformer (ViT) [8] to extract feature for each image $\mathbf{i}^t \in \mathbb{R}^{H \times W \times 3}$ separately. A ViT has two main steps: image patch embedding generation and transformer encoding. In the first step, an image is split into N^p fixed-sized square patches, where $N^p = (W/P) \times (H/P)$ and P is the size of a patch. Then each patch is flattened and linearly projected into a latent space by the image patch embedding layer. An extra learnable embedding is prepended as a token for classification (called "CLS") to the patch embeddings. Besides, position embeddings are added to the patch embeddings to retain positional information. In the second step, the resulting patch-embedding sequence, represented by $\mathbf{E}^i \in \mathbb{R}^{(H \times W + 1) \times D}$ and D denotes dimension of embeddings, is fed to the ViT encoder. A ViT encoder consists of N^e stacked ViT encoder blocks, and each block contains a Multi-Head Self-Attention (MHSA) layer and a Position-wise Feed-Forward Network (PFFN). Both MHSA and PFFN have a layer normalization and a residual connection applied to them. The output of ViT encoder is image feature $\mathbf{F}^i \in \mathbb{R}^{(H \times W + 1) \times D}$, which consists of a sequence of patch features and a global classification feature.

Then, we concatenate the features of all images in \mathbf{I}^t and use extra N^s stacked transformer encoder blocks [33] to establish the relationship between features from different images. The output of these extra encoder blocks is sequence feature $\mathbf{F}^e \in \mathbb{R}^{(H \times W + 1) K^t \times D}$, which is also the output of MUIsC encoder.

3.4.2 Decoder. The decoder is an autoregressive Natural Language Generation (NLG) model with vision-language fusion and an additional classification head. Our decoder is made up of a stack of N^d

decoder blocks, and each decoder block consists of three layers: a masked MHSA, a Multi-Head Cross-Attention (MHCA) and a PFFN. Each of these three layers is followed by a residual connection and a layer normalization. Masked MHSA has a similar architecture with the MHSA in MUIsC encoder. But different from MHSA that is applied to all tokens (i.e. bi-directional attention mechanism), masked MHSA only collects information from the prior tokens (i.e. uni-directional attention mechanism). MHCA is also similar with MHSA except that MHCA takes two sequences of embeddings/features as inputs, and then computes the relationship between them. In our method, MHCA's two inputs include the output from MUIsC encoder, i.e. \mathbf{F}^e and the embedding sequence from the previous masked MHSA. Note that MHCA may be absent in some decoder blocks to keep our model concise. The input of the masked MHSA of the first decoder block is a sequence of K^d word embeddings, $\mathbf{E}^d = \{\mathbf{e}^d_1, \dots, \mathbf{e}^d_{K^d}\}$, which are generated from a sequence of words, denoted as \mathbf{W}^d . In the training stage, \mathbf{W}^d is constructed by concatenating product title \mathbf{W}^t and textual feedback \mathbf{W}^f separated by a special token s , i.e. $\mathbf{W}^d = \{w_1^t, \dots, w_{K^T}^t, s, w_1^f, \dots, w_{K^f}^f\}$, where \mathbf{W}^t provides additional textual information for a product and \mathbf{W}^f serves as the groundtruth in the training of autoregressive NLG model. In the inference stage, \mathbf{W}^d only contains \mathbf{W}^t and a following s . We tokenize \mathbf{W}^d into a sequence of tokens and encode the resulting tokens to word embeddings \mathbf{E}^d via a word embedding layer. Then, position embeddings are added to \mathbf{E}^d . Taking word embeddings \mathbf{E}^d and image-sequence feature \mathbf{F}^e as input, our decoder finally outputs a sequence of decoded hidden states \mathbf{H}^d .

	Qualified	Single-image	Image-pair	Multi-image
Prop.	0.71	0.12	0.07	0.10

Table 1: Proportions of qualified samples, and samples rejected due to single-image, image-pair, and multi-image rule violations in AGPIS-data.

3.4.3 Multi-Task Learning. We combine two tasks in a single MUIsC model - NLG which generates a textual feedback, and sequence Multi-class Classification (McC) which classifies a sequence to a qualified one or a category of rule violations.

Given an input word sequence which consists of product title \mathbf{W}^t paired with review feedback \mathbf{W}^f and a sequence of images \mathbf{I}^t , our NLG task is to estimate the conditional probability for each token in \mathbf{W}^f :

$$P^{nlg}(\mathbf{W}^f | \mathbf{I}^t, \mathbf{W}^t) = \prod_{q=1}^{K^f} p^{nlg}(w_q^f | w_{<q}^f; \mathbf{I}^t, \mathbf{W}^t) \quad (2)$$

where $w_{<q}^f$ stands for all tokens prior to position q (i.e. $w_{<q}^f = (w_1^f, \dots, w_{q-1}^f)$). In the training stage, \mathbf{W}^t is located prior to \mathbf{W}^f , and thus the $p^{nlg}(w_q^f)$ is conditioned on the preceding tokens in $\mathbf{W}^f, \mathbf{I}^t$, and all tokens in \mathbf{W}^t .

Given the whole training set $\mathcal{N} = (\mathbf{W}^t, \mathbf{I}^t, \mathbf{W}^f)$, our NLG task can be trained by optimizing the loss function as follows:

$$\mathcal{L}^{nlg}(\mathcal{N}) = - \sum_{(\mathbf{W}^f, \mathbf{I}^t, \mathbf{W}^t) \in \mathcal{N}} \log P^{nlg}(\mathbf{W}^f | \mathbf{I}^t, \mathbf{W}^t). \quad (3)$$

If we set textual feedback of a qualified image-sequence to a constant word, e.g. "yes", i.e. $K^f = 1$ and $w_1^f = \text{"yes"}$, $p^{nlg}(w_1^f)$ is actually the probability of \mathbf{I}^t being qualified.

Sequence multi-class classification task classifies a sequence of images \mathbf{I}^t into K^g classes $\mathbf{G} \in \{g_1, \dots, g_{K^g}\}$, where g_1 is the class of being qualified and $\{g_2, \dots, g_{K^g}\}$ represent the classes of being rejected due to various rule violations. A McC head is applied on the decoded hidden state of token s which follows \mathbf{W}^t , and thus our McC task is conditioned on \mathbf{W}^t and \mathbf{I}^t . We use a softmax classifier where the loss function is

$$\mathcal{L}^{mcc}(\mathcal{N}) = \sum_{(\mathbf{W}^f, \mathbf{I}^t, \mathbf{W}^t) \in \mathcal{N}} \left(- \sum_{i=1}^{K^g} 1(c(\mathbf{I}^t) = i) \log(p_i^{mcc}(\mathbf{I}^t)) \right)$$

, where $1(\cdot)$ is the indicator function, $c(\mathbf{I}^t)$ is the class label of \mathbf{I}^t , and $p_i^{mcc}(\mathbf{I}^t)$ represents the probability of \mathbf{I}^t belonging to class i . Especially, $p_1^{mcc}(\mathbf{I}^t)$ is the probability of \mathbf{I}^t being qualified.

Combining the natural language generation and multi-class classification, the loss function for MUIsC is

$$\mathcal{L}(\mathcal{N}) = \lambda^{nlg} \mathcal{L}^{nlg}(\mathcal{N}) + \lambda^{mcc} \mathcal{L}^{mcc}(\mathcal{N}) \quad (4)$$

, where $\lambda^{nlg}, \lambda^{mcc} > 0$ are factors that balance the two loss functions. In the inference stage, our framework takes $p_1^{mcc}(\mathbf{I}^t)$ as its output, i.e. $p^t = p_1^{mcc}(\mathbf{I}^t)$.

4 OFFLINE EVALUATION OF MUISC

4.1 Dataset and Metrics

The stage 1 of our proposed framework is firstly deployed to produce image sequences for JD.com's "Haohuo" channel. With months of data accumulation, we collect produced three-image sequences and corresponding textual feedback as our dataset for MUIsC training and evaluation. Besides, textual product title and candidate images are also collected from JD.com's main site for each product in the dataset. In this paper, we use a dataset collected within three months and call it as AGPIS-data. AGPIS-data contains over 700K samples from 39 product categories. Distribution details of this dataset can be found in Table 1. We can see that about 29% samples in our dataset are rejected ones. Note that, because part of image files are not valid anymore when we collect them and we only keep the samples with valid images, the proportion of qualified samples in our dataset is not equivalent to the acceptance rate in real production.

We consider the rule name appeared in a textual feedback as the class label of a sample for the learning of multi-class classification task. If there are more than one rules in a textual feedback, we just choose the first one. In our dataset, only 4.7% rejected samples have more than one violated rules. There are mainly 43 image-relevant rules in our AGPIS-data, and our multi-class labels have 45 classes with one extra class for qualified samples and another class for the samples rejected by other image-relevant rules. We can also convert multi-class labels to binary-class ones by simply merging 44 classes of rejected samples into one class. AGPIS-data is randomly split into three subsets without any overlap in product SKUs – training (80%), validation(10%), and testing(10%). Besides, in order to evaluate the detection performance of different categories of rule violations, we also build datasets AGPIS-data-single, AGPIS-data-pair, and AGPIS-data-multi for single-image rules, image-pair rules, and multi-image rules, respectively. Each of the above datasets contains the same number of randomly selected qualified samples and samples rejected by a specific category of rule violations.

ROC AUC (AUC) and Recall@Precision (R@P) are used to evaluate the performance of models. AUC is defined as the area under the ROC curve and R@P is the recall value at a given precision.

4.2 Implementation details

MUIsC is trained in an end-to-end manner with the encoder initialized by a pretrained ViT model and the decoder initialized by a pretrained GPT2 [27] model. We use a pretrained ViT-B16-224 ($N^e = 12$), the base version of the ViT with ($P = 16$) patch size and ($H \times W = 224 \times 224$) input image size. This model is pretrained on ImageNet-21k and finetuned on ImageNet 2012. For decoder, we use a GPT2 model pretrained on CLUECorpus [35], a large-scale Chinese Corpus. Our decoder only has $N^d = 3$ blocks to keep the model concise and efficient, since our textual product title and review feedback are relatively simple. Embedding dimension in both encoder and decoder is $D = 768$. The model is trained for 10 epochs using the AdamW optimizer, with batch size 64. We use 1.5e-4 as the initial learning rate in our experiments and real applications.

Other parameters that need to be set in the proposed method are

- Number of images in target image sequence is $K^t = 3$, and number of candidate images is $K^c \approx 7$.

Architecture	Backbone	Fusion mode	AUC	R@P=0.8	R@P=0.85	R@P=0.9	AUC-single	AUC-pair	AUC-multi
Single-Tower	ResNet18	early	0.745	0.803	0.592	0.344	0.739	0.736	0.737
		late	0.755	0.834	0.616	0.369	0.780	0.728	0.737
	ResNet50	early	0.751	0.806	0.620	0.415	0.746	0.735	0.744
		late	0.763	0.841	0.647	0.392	0.792	0.724	0.743
	ResNetV2-101(BiT)	early	0.765	0.843	0.639	0.444	0.763	0.775	0.752
		late	0.767	0.840	0.667	0.442	0.788	0.726	0.749
	ViT	early	0.763	0.849	0.631	0.412	0.749	0.784	0.751
		late	0.764	0.831	0.644	0.429	0.772	0.729	0.767
	ViT [*]	late	0.758	0.830	0.634	0.382	0.762	0.733	0.752
Ours	ViT	late	0.800	0.904	0.758	0.524	0.795	0.774	0.791

Table 2: Performance comparison. Baselines use different visual backbones and image fusion methods. Superscript * indicates that the model is trained on a binary-class task and the other models are trained on a multi-class task.

- Number of extra encoder blocks for image-feature interaction is $N^s = 1$.
- Balance factor in the loss function in Eq. 4 is $\lambda^{nlg} = 0.1$ and $\lambda^{mcc} = 1.0$.

4.3 Baseline Methods

Our MUIsC aims to solve the problem of binary image-sequence classification, which can be considered as an extension of single-image classification. Note that a specific issue in image-sequence classification is how to fuse the information of multiple images in a sequence. To validate MUIsC, we build baselines based on the classic single-image classification architecture, called single-tower, which consists of a visual backbone and a classification head. We experiment with two image-fusion methods (early fusion and late fusion) and 4 backbones (ResNet18[11], ResNet50[11], ResNetV2-101[18], and the ViT[8] used in MUIsC). The early fusion method concatenates all images of a sequence into a single image, while the late fusion method firstly extracts features for each image and then concatenates these features together. The classification head is a MLP classifier. Models are all trained on AGPIS-data.

4.4 Performance Comparison

Table2 shows the performance of our method and baselines on different datasets, where AUC, AUC-single, AUC-pair, and AUC-multi represent the AUC on AGPIS-data, AGPIS-data-single, AGPIS-data-pair, and AGPIS-data-multi, respectively. Besides the model with superscript *, all models are trained on a multi-classification task. We can observe that the proposed MUIsC outperforms other methods in AUC on AGPIS-data and achieves the best results in terms of all the other evaluation metrics and datasets except for AUC on AGPIS-data-pair.

Among all visual backbones, ResNetV2-101 and ViT performs better than ResNet18 and ResNet50. This shows that powerful backbones are able to play some role in AGPIS. The comparison between early and late fusion modes indicates that each mode has its own strength and weakness, though late-fusion methods outperform early-fusion ones on the whole dataset. Early-fusion methods perform better on AGPIS-data-pair dataset. We think this is caused by the deep interaction between images. Late-fusion methods achieve

better on AGPIS-data-single dataset since image features are extracted individually in the early stage. Our MUIsC adopts late-fusion considering the overall performance. But it is still an interesting topic to balance model performance between single-image rules and image-pair rules. Besides, early-fusion and late-fusion methods have similar performance on AGPIS-data-multi. We speculate the reason is that AGPIS-data-multi emphasizes both single-image features and interaction between images. We also show the performance of the model trained on binary classification task, and observe that binary-class classifier performs worse than its corresponding multi-class one. This indicates that more detailed information about rejection can help improve model performance. In MUIsC, we further use textual feedback to introduce more information.

4.5 Ablation Study

The structure of MUIsC is carefully ablated with the results listed in Table 3. Here, we use the same set of datasets and metrics with Section 4.4. The baseline for ablation study is a single-tower model trained on McC task, which either does not have decoder or uses any textual data. By introducing hierarchical image feature extraction method, a 1.4% AUC gain is achieved, which shows the extra N^s transformer encoder blocks result in better interaction between images than baseline late fusion. After using an encoder-decoder architecture, the performance is slightly improved, and indicates that directly adding a decoder does not lead to a big performance gain. Then, we add the NLG task and formulate a multi-task learning model. It is observed that the performance is considerable improved, which indicates that the textual review feedback effectively guides the model to better understand rejected samples during training and results in better performance. By including textual product title as an extra input, we get a AUC gain of 0.8% and the best performance is achieved. Interestingly, two kinds of textual information play different roles in terms of performance improvement. The NLG task on textual feedback leads to a good AUC gain on AGPIS-data-single, while product-title input achieves a significant gain on AGPIS-data-multi. This verifies that the detections of different categories of rule violations require different information for AGPIS.

Single tower	Hier.	Enc.-dec.	McC task	NLG task	Text as input	AUC	R@P=0.8	R@P=0.85	R@P=0.9	AUC-single	AUC-pair	AUC-multi
✓			✓			0.764	0.831	0.644	0.429	0.772	0.729	0.767
✓	✓		✓			0.778	0.856	0.701	0.456	0.781	0.755	0.762
✓	✓	✓	✓			0.780	0.876	0.688	0.450	0.784	0.757	0.762
✓	✓	✓	✓	✓		0.792	0.891	0.730	0.507	0.799	0.764	0.765
✓	✓	✓	✓	✓	✓	0.800	0.904	0.758	0.524	0.795	0.774	0.791

Table 3: Ablation of components in MUIsC.



Figure 5: Examples of qualitative AGPIS framework results.

4.6 Qualitative Analysis

Four qualitative example results are illustrated in Figure 5. Each example includes a "good" sequence and a "bad" one for a same product, according to the probability of being qualified p^t estimated by our framework (shown above each sequence). We can observe that all the sequences with low p^t (the second row) violates our rules. The sequence in the leftmost example contains non-compliant banner and logo in the primary image. For the second left example, the primary and the third images contain products with different colors, which violates an image-pair rule. Besides, the second right and rightmost examples have improper display orders since the primary image fails to provide a whole picture of the product and violates multi-image rules. Meanwhile, we can see that all good sequences (the first row) have compliant single images and present the product in a proper order. These examples show that the proposed framework is effective in detecting all categories of rule violations and generating qualified image sequences.

5 ONLINE EVALUATION

5.1 Deployment and Online Evaluation

JD "Haohuo" Channel (Discovery Goods Channel) is an important traffic entrance for JD.com and also a good platform for users to discover their potential purchase interests, thus the product quality and presentation is of great influence for platform's income. Before the release of our AGPIS framework, image selection for products mainly depends on human labors, which may be expensive and perform lower efficient. Since Feb 2021, our AGPIS framework has been deployed on "Haohuo" Channel (both website and App), and has produced high-standard images for about 1.5 million featured products. The amount of production is equivalent to 1000+ human professions.

Our framework is deployed step by step. Stage 1 is developed and deployed firstly. Then, stage 2 is trained on the resulting data

and is deployed following stage 1. Our MUIsC model is updated every 3 months and is trained on AGPIS-data collected within the past 3 months. We take reject rate by human reviewers as online evaluation metric and set the submission threshold th^t to 0.3. According to the statistical results, the reject rate is 19.3% for the period when there is only stage 1, and it is further reduced to 13.6% after stage 2 is added. Note that in order to avoid high reject rate in the period when stage 2 is not ready, stage 1 works under a strict setting, which may lead to high false-negative. In the future, we will gradually relax stage 1 and enable stage 2 to select among multiple candidates instead of just evaluating one candidate.

5.2 System A/B testing

In "Haohuo" Channel, an image-text system is built with our AGPIS framework and a product-copy generation framework[10][39] to generate both images and textual description for a product. To evaluate the effectiveness of this system, we compare key online metrics before and after deploying our system on "Haohuo" Channel. The business of the platform is measured by Click-Through Rate (CTR) and Conversion Rate (CVR). The A/B testing is conducted on most of the main categories of products, covering clothes, electronics, computers, beauty & health, groceries, etc. For users in baseline group, the platform shows them product images and text generated by human professions. For users in experiment group, the platform shows them images and text generated by our system. We find that images and text generated by our system outperform the ones submitted by human professions. Specifically, our system improves CTR by 3.2% and CVR by 3.6% over baseline. The increase of CTR indicates that the users prefer to click products with product images and text generated by our system. The improvement of CVR demonstrates that the images and text generated by our system is successful in convincing users to make purchase decisions.

Currently, the product image-text generation task in JD "Haohuo" Channel is jointly performed by human professions and our

image-text system. We observe a trend that human professions sometimes prefer to use our system to assist their generation process. In addition, our system can also benefit for long-tail products, making the channel a healthier ecosystem.

6 CONCLUSION

In this paper, we developed a framework for Automatic Generation of Product-Image Sequence (AGPIS) in e-commerce. To address the unique challenges in AGPIS, our framework is designed as a combination of rule-based specific methods (stage 1) and a Multi-modality Unified Image-sequence Classifier (MUIsC) (stage 2) which is able to detect all categories of rule violations while considering multi-modality information. The experimental results show that our MUIsC outperforms various baselines. Our framework has been deployed on JD.com's "Haohuo" Channel, and a high volume of 1.5 million product image sequences have been generated. With the help of our AGPIS framework, our CTR is improved by 3.2% and CVR is improved by 3.6%.

REFERENCES

- [1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.
- [2] Abon Chaudhuri, Paolo Messina, Samrat Kokkula, Aditya Subramanian, Abhinandan Krishnan, Shreyansh Gandhi, Alessandro Magnani, and Venkatesh Kandaswamy. 2018. A smart system for selection of optimal product images in e-commerce. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 1728–1736.
- [3] Ming-Yi Chen and Ching-I Teng. 2013. A comprehensive model of the effects of online store image on purchase intention in an e-commerce environment. *Electronic Commerce Research* 13, 1 (2013), 1–23.
- [4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*. Springer, 104–120.
- [5] Ritendra Datta, Dhairaj Joshi, Jia Li, and James Z Wang. 2006. Studying aesthetics in photographic images using a computational approach. In *European conference on computer vision*. Springer, 288–301.
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [7] Wei Di, Neel Sundaresan, Robinson Piramuthu, and Anurag Bhardwaj. 2014. Is a picture really worth a thousand words? -on the role of images in e-commerce. In *Proceedings of the 7th ACM international conference on Web search and data mining*. 633–642.
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11292* (2020).
- [9] Shreyansh Gandhi, Samrat Kokkula, Abon Chaudhuri, Alessandro Magnani, Theban Stanley, Behzad Ahmadi, Venkatesh Kandaswamy, Omer Ovenc, and Shie Mannor. 2019. Image matters: Detecting offensive and non-compliant content/logo in product images. *arXiv preprint arXiv:1905.02234* (2019).
- [10] Xiaojie Guo, Shugen Wang, Hanqing Zhao, Shiliang Diao, Jiajia Chen, Zhuoye Ding, Zhen He, Yun Xiao, Bo Long, Han Yu, et al. 2021. Intelligent Online Selling Point Extraction for E-Commerce Recommendation. *arXiv preprint arXiv:2112.10613* (2021).
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [12] Ronghang Hu and Amanpreet Singh. 2021. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1439–1449.
- [13] Bin Jin, Maria V Ortiz Segovia, and Sabine Süsstrunk. 2016. Image aesthetic predictors based on weighted CNNs. In *2016 IEEE International Conference on Image Processing (ICIP)*. Ieee, 2291–2295.
- [14] Alexis Joly and Olivier Buisson. 2009. Logo retrieval with a contrario visual query expansion. In *Proceedings of the 17th ACM international conference on Multimedia*. 581–584.
- [15] Michael J Jones and James M Rehg. 2002. Statistical color models with application to skin detection. *International journal of computer vision* 46, 1 (2002), 81–96.
- [16] Le Kang, Peng Ye, Yi Li, and David Doermann. 2014. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1733–1740.
- [17] Yan Ke, Xiaou Tang, and Feng Jing. 2006. The design of high-level features for photo quality assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 1. IEEE, 419–426.
- [18] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. 2020. Big transfer (bit): General visual representation learning. In *European conference on computer vision*. Springer, 491–507.
- [19] Dmitry Kuzovkin, Tania Pouli, Olivier Le Meur, Rémi Cozot, Jonathan Kerec, and Kadi Bouatouch. 2019. Context in photo albums: Understanding and modeling user behavior in clustering and selection. *ACM Transactions on Applied Perception (TAP)* 16, 2 (2019), 1–20.
- [20] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.
- [21] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).
- [22] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. 2021. Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804* (2021).
- [23] Yinhai Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
- [24] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265* (2019).
- [25] Eftichia Mavridaki and Vasileios Mezaris. 2015. A comprehensive aesthetic quality assessment method for natural images using basic rules of photography. In *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 887–891.
- [26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020* (2021).
- [27] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
- [28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).
- [29] Stefan Romberg and Rainer Lienhart. 2013. Bundle min-hashing for logo recognition. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. 113–120.
- [30] Kekai Sheng, Weiming Dong, Menglei Chai, Guohui Wang, Peng Zhou, Feiyue Huang, Bao-Gang Hu, Rongrong Ji, and Chongyang Ma. 2020. Revisiting image aesthetic assessment via self-supervised feature learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5709–5716.
- [31] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).
- [32] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. 2013. Selective search for object recognition. *International journal of computer vision* 104, 2 (2013), 154–171.
- [33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
- [34] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.
- [35] Liang Xu, Xuanwei Zhang, and Qianqian Dong. 2020. CLUECorpus2020: A large-scale Chinese corpus for pre-training language model. *arXiv preprint arXiv:2003.01355* (2020).
- [36] Haiming Yin, Xiaodong Xu, and Lihua Ye. 2011. Big skin regions detection for adult image identification. In *2011 Workshop on Digital Media and Digital Content Management*. IEEE, 242–247.
- [37] Wenhui Yu, Huidi Zhang, Xiangnan He, Xu Chen, Li Xiong, and Zheng Qin. 2018. Aesthetic-based clothing recommendation. In *Proceedings of the 2018 world wide web conference*. 649–658.
- [38] Stephen Zakrewski, Kamelia Aryafar, and Ali Shokoufandeh. 2016. Item popularity prediction in e-commerce using image quality feature vectors. *arXiv preprint arXiv:1605.03663* (2016).
- [39] Xueying Zhang, Yanyan Zou, Hainan Zhang, Jing Zhou, Shiliang Diao, Jiajia Chen, Zhuoye Ding, Zhen He, Xueqi He, Yun Xiao, et al. 2021. Automatic Product Copywriting for E-Commerce. *arXiv preprint arXiv:2112.11915* (2021).
- [40] C Lawrence Zitnick and Piotr Dollár. 2014. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*. Springer, 391–405.