

Received October 12, 2021, accepted November 4, 2021, date of publication November 10, 2021, date of current version November 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3127324

# Automatic Chinese Meme Generation Using Deep Neural Networks

LIN WANG<sup>1</sup>, QIMENG ZHANG<sup>1</sup>, YOUNGBIN KIM<sup>2</sup>, (Member, IEEE), RUIZHENG WU<sup>3</sup>, HONGYU JIN<sup>1</sup>, HAOKE DENG<sup>1</sup>, PENGCHU LUO<sup>1</sup>, AND CHANG-HUN KIM<sup>1</sup>

<sup>1</sup>Department of Computer Science and Engineering, Korea University, Seoul 02841, South Korea

<sup>2</sup>Graduate School of Advanced Imaging Science, Multimedia and Film, Chung-Ang University, Seoul 06974, South Korea

<sup>3</sup>Department of Computer Science and Engineering, The Chinese University of Hong Kong, Central Ave, Hong Kong

Corresponding author: Chang-Hun Kim (chkim@korea.ac.kr)

This work was supported by 10.13039/501100014188-Korea Government [Ministry of Science and ICT (MSIT)] under Grant NRF-2021R1A2C1094624.

**ABSTRACT** Internet memes have become widely used by people for online communication and interaction, particularly through social media. Interest in meme-generation research has been increasing rapidly. In this study, we address the problem of meme generation as an image captioning task, which uses an encoder–decoder architecture to generate Chinese meme texts that match image content. First, to train the model on the characteristics of Chinese memes, we collected a dataset of 3,000 meme images with 30,000 corresponding humorous Chinese meme texts. Second, we introduced a Chinese meme generation system that can generate humorous and relevant texts from any given image. Our system used a pre-trained ResNet-50 for image feature extraction and a state-of-the-art transformer-based GPT-2 model to generate Chinese meme texts. Finally, we combined the generated text and images to form common image memes. We performed qualitative evaluations of the generated Chinese meme texts through different user studies. The evaluation results revealed that the Chinese memes generated by our model were indistinguishable from real ones.

**INDEX TERMS** Deep learning, computer vision, image captioning, meme generation, internet meme.

## I. INTRODUCTION


The sharing of memes is a widely adopted social phenomenon used to express thoughts and opinions in a creative or humorous manner [1], particularly on social media. Memes are typically comprised of screenshots of celebrities, popular animations, movies, or personal images as sources, accompanied by humorous text that matches the images with interesting and subtle connotations. In recent years, memes have been widely used in applications such as Twitter [2] and Facebook [3]. In addition, they are also widely used in Chinese SNSs (e.g., Sina Weibo [4], WeChat [5]). One reason for why memes are widely spread is that they can express more abstract emotions to compensate for the lack of chatting language in an easy and amusing manner.

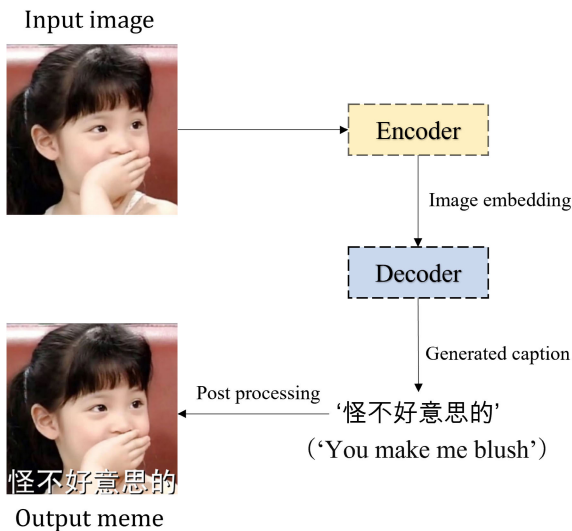
Although memes are widely used in social media, most are created manually. That is, people observe image content, come up with related humorous text that can express the meanings of the images first, and then manually paste the text onto the image using image editing tools to complete the

meme. It requires people to have a sense of humor, while the process of image editing is also time-consuming. Therefore, an automatic method for generating memes is necessary and helpful.

In recent years, learning-based research in the fields of computer vision [6] and natural language processing [7] have been very active, such as in image classification [8], [9], object detection [10], [11], machine translation [12], [13], and language understanding [14], [15]. They all have promoted research on multimodal learning in these two fields, such as in image captioning [16]–[18] and visual question answering [19], [20]. They have also promoted research on automatic meme-generation tasks. These also facilitate the automatic generation of memes. Recently, several studies [21], [22] have examined the automatic generation of memes and successfully made meme generation websites or applications [23], [24]. However, these previous studies all focused on English meme generation, and research on the automatic generation of Chinese memes is scarce.

This study aims to generate Chinese memes according to a given single image. Based on the process of manual meme creation, as illustrated in Figure 1, we define the Chinese

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Zhao .



**FIGURE 1.** Illustrative figure of our Chinese meme generation process.

meme generation task in three steps: 1) extract features of the input image; 2) use these image features to generate relevant Chinese text that is humorous and can express the meaning of the input image; 3) combine the generated text with the input image to complete the final image memes. Inspired by image captioning methods [11], [25], we modeled the process of extracting features from images and generating text into an encoder–decoder architecture, which consists of a convolutional neural network that extracts features from input images, combined with a language model to generate text using image features.

To achieve this task, we first constructed a dataset of 3,000 meme images from Google Images and other Chinese meme collection websites, as there is currently no public dataset of Chinese image memes. Each image corresponds to 10 meme captions. Then, we employed ResNet-50 [26] combined with the OpenAI GPT-2 [27] model to generate Chinese meme texts. Finally, we followed the general format of Chinese memes and pasted the text to the bottom of the input image to obtain the final Chinese image meme. Further, we provided an in-depth qualitative analysis based on user studies and obtained positive evaluation results.

This study makes the following contributions:

- We present a Chinese meme generation system based on the image captioning model architecture to generate Chinese memes creatively.
- We constructed a Chinese meme dataset for the first time, which contains 3,000 meme images with each image corresponding to 10 Chinese meme sentences.
- Through qualitative analysis of user studies, we prove that the proposed method can generate Chinese memes that cannot be distinguished from the manually created ones.

## II. RELATED WORK

Although Internet memes are widely used, there are only a few studies on automatic image meme generation.

Meme-generation studies can be divided into two categories. One automatically generates memes based on text input, and the other automatically generates memes based on the image input. However, most of these studies only focused on the generation of English memes.

### A. ENGLISH MEME GENERATION

For English meme generation, there are some studies based on text input as well as image input. For text-input meme generation, Sadasivam *et al.* [28] regarded meme generation as a process of text translation. They extracted features from the input text and used a selection module to select a meme template image from a set of popular candidate images based on the text features, generated text from the selected template image, and then combined the meme template image and the generated text. For image-input meme generation, Wang and Wen [29] proposed a meme-text sorting algorithm based on input image features and caption candidate features. They first performed a reverse image search on the web for the input image to obtain the best guess for the image as the keyword of the sentence to be generated. Then, they searched for other relevant caption candidates based on keywords on meme-searching websites. Finally, they used input image features to sort these candidates according to the proposed sorting algorithm to select the best caption generation result as the caption output for the input image. Vyalla and Udandarao [22] proposed an end-to-end meme generation system that allows users to generate memes on their web applications in real time. They trained a classification model to classify the input image according to their own class labeling of images to obtain the class label, and then they forced the generation model to generate sentences from the label information. Although the above-mentioned studies could generate English image memes, they did not directly generate meme images from the input sentences or images, so these indirect generation methods may have more losses.

Peirson and Tolunay [21] regarded the meme generation task as an image captioning task and proposed a model that could directly use the input image features to generate sentences. First, they created an English meme dataset of 40,000 images, each containing approximately 160 captions, and they used a simple encoder–decoder architecture, which was proposed in [16] with an attention mechanism [30], to generate meme sentences from input images. Instead of selecting candidate images or sentences from the existing data, such as the above-mentioned studies, they directly used the obtained image features to generate captions so that the losses in the process may be smaller than those of the other methods mentioned above. They generated meme text, broke the sentence into upper and lower lines, and placed them on the image manually according to the general format of the English memes. However, one shortcoming of this study is that they used the earliest image captioning model [16], which was proposed in 2015 as their baseline model; they did not use the models with higher performance that appeared later.

## B. CHINESE MEME GENERATION

For Chinese meme generation, there is one study [31] on the generation of image memes, but only panda face memes were generated from input sentences. They used a generative adversarial network [32] with an attention module and template panda face information as supplementary signals to add or change some details to align the semantics of the template panda face image to the input sentence. However, this study is limited to generating memes of panda face images, which are often used in China. Some reasons why Chinese memes are widely used but there is a lack of research may be that the Chinese meme dataset is difficult to collect, the quality evaluation criteria for generated memes is difficult to define, and Chinese expressions are more complicated and diverse.

## C. IMAGE CAPTIONING

As mentioned above, the process of Chinese meme sentence generation defined in this study can be implemented using image captioning models. The problem of generating natural language descriptions from images has long been studied in the field of computer vision. Image captioning requires models to not only be able to detect objects in the image, but also understand the relationship between the objects, and finally express using reasonable phrasing related to the image. Traditional image captioning models are based on an encoder–decoder architecture [33]. It converts the input image into a fixed-length vector and then converts this vector into an indefinite-length output sequence. Various application models have been designed based on the encoder–decoder architecture. Vinyals *et al.* [16] created the first widely recognized model for generating image descriptions. They used a deep convolutional neural network [34] to extract a fixed-length visual feature vector from the input image and then connected a recurrent network [35] as a decoder to convert this feature vector into an output text sequence to model the process from the input image to text generation.

Although the encoder–decoder architecture is very effective, it is also very limited. Encoders and decoders are only connected by a fixed-length feature vector, but this feature vector may not be sufficient to express the entire input, which may make it impossible to obtain sufficient information from the input at the beginning of the decoding process, so the accuracy of decoding may not be very high. Therefore, to compensate for the deficiencies of the basic encoder–decoder models, the attention mechanism [36] was proposed and applied in the natural language processing field. It is used not only to pay attention to the global semantic coding vector when generating words, but also to add an “area of interest”, which indicates the parts of the input sequence that should be focused on when generating the subsequent words.

Referring to this attention mechanism, Xu *et al.* [25] first introduced attention to image captioning models. They added attention weights to each position of the feature maps of the convolutional layers. These weights represent the attention

factor, which can be learned through back propagation. Then, the weighted feature vector is input to the decoder. Using the attention mechanism, the model can determine where focus on in the image when generating the current word, making sentence generation from input images more reasonable. Since then, the attention mechanism has been widely used in the field of image captioning [25], [37], [38]. The encoder typically uses a deep convolutional neural network [26], [39], which can generate embedded fixed-length image feature vectors from the input image. For the decoder, the long short-term memory (LSTM) [35] recurrent networks can solve the long sequence dependence problem, so they have excellent performance in sequence generation tasks and have been widely used until the Transformer [40] was proposed. The Transformer abandoned the traditional complicated convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which rely entirely on the attention mechanism to capture the long-distance dependence in the sequence, and it can process multiple data in parallel, which greatly reduces the training time. The Transformer encoder architecture is widely used to convert words into meaningful vectors, such as Google Bert [41], and the Transformer decoder or the full Transformer architecture is used to generate new words from other words such as OpenAI GPT [42]. Among them, the OpenAI GPT-2 [27] exhibited an impressive ability to write coherent and passionate essays that exceeded what we anticipated from current language models. Therefore, to generate more fluent and emotional sentences, we hope to use OpenAI GPT-2 as a decoder.

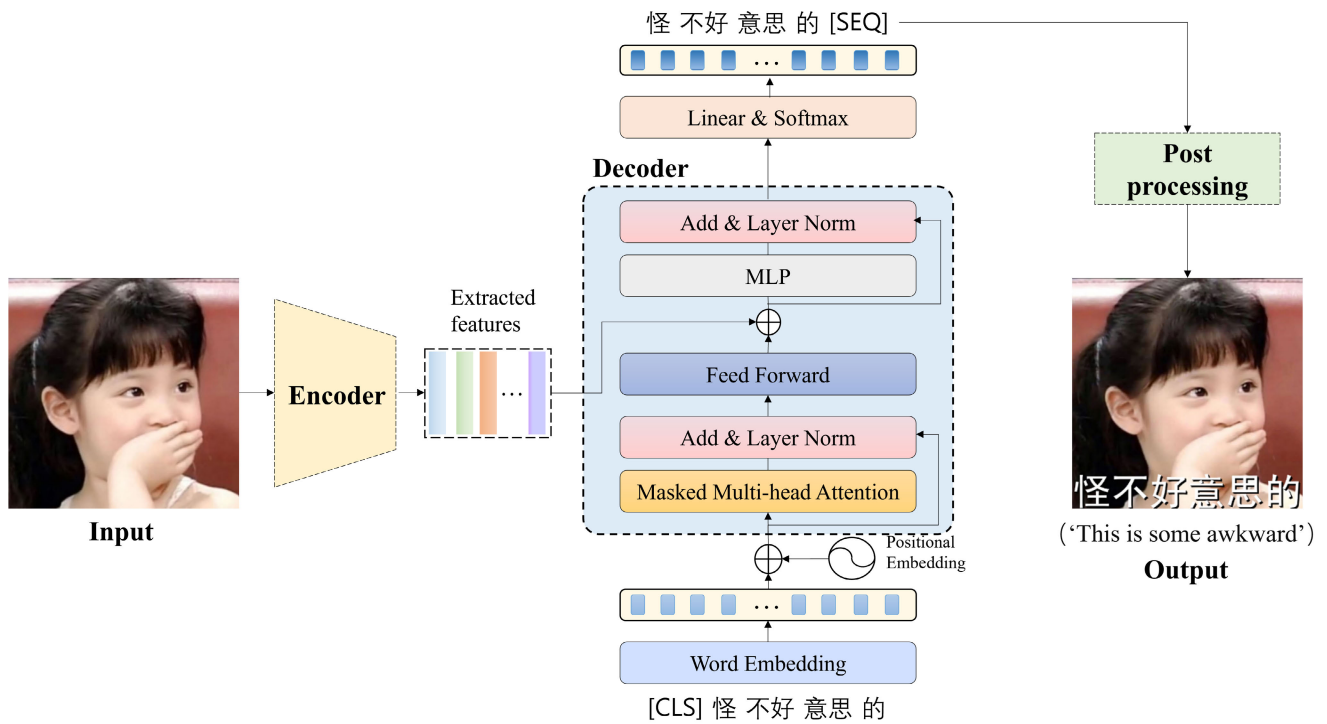
In this paper, inspired by Peirson and Tolunay [21], we also regarded Chinese image meme generation as an image captioning problem and employed an encoder–decoder model to solve it. Input images were used to generate the Chinese image memes directly. In the following section, we describe the proposed method in detail.

## III. METHOD

In this section, we first introduce an overview of our Chinese meme generation system and then explore the dataset we constructed for the Chinese image meme generation task. Finally, we explore the process of generating Chinese image memes through the model in detail.

### A. OVERVIEW

In this task, our goal is to generate Chinese image memes that are related to the input image content and express the humor of the images. We first created a dataset of Chinese meme images and collected Chinese captions for them as training data in Section III-B. Second, we developed an image captioning model using a CNN as the base image feature extraction architecture to encode the input image into a fixed-length feature vector and connected it with a Transformer architecture to generate Chinese captions using the input image feature vector obtained, which is shown in Section III-C. Finally, we visualized the generated Chinese captions on the input images by using the relationship between the size



**FIGURE 2.** Architecture of the proposed Chinese meme-generation system. As shown in the figure, the model contains two components: the encoder is a CNN to extract the features of the input image, and the decoder is a transformer decoder-based model to generate Chinese captions by using the image features obtained. After the model, a post-processing operation combines the generated text with the input image to complete the Chinese image meme generation.

of the images and the generated texts to place the captions to the bottom of the images and obtain the final Chinese meme images in Section III-D. The network architecture and workflow are illustrated in Figure 2.

**B. DATASET**

The general image captioning task requires the use of image features to generate captions that can describe the content of the input image. However, the meme generation task should meet the requirements that the generated caption must be relevant to the image and that the generated captions should be humorous. Especially for deep learning models, sufficient training data is necessary to drive a model to learn the features of Chinese meme images and the characteristics of Chinese sentences. However, contemporary meme-related datasets available online have various limitations. The Reddit meme dataset on kaggle [43] and the meme generator dataset [44] are both English meme datasets, and they only have memes with texts on the images, and they cannot distinguish between the images and the text. Currently, Chinese meme datasets are not available. In this paper, for the first time, we present a Chinese image meme dataset.

To build our dataset, first, we collected 3,000 Chinese meme images from Google Images and some Chinese meme production websites that provide various meme templates and allow users to upload images and edit text, such as ‘52doutu’ [45]. Second, we extracted the text on these meme

images as our sentence data. However, according to the characteristics of memes, different people may have different interpretations of the same image, and so, the same image may be paired with different sentences. Therefore, we had to match various text data to each meme image. To meet the requirement of matching multiple sentences with correct expression for each image, and at the same time, eliminate the subjectivity of data selection, we collected text data using the following steps:

- 1) Collect Chinese meme images on Google Images and some Chinese meme websites.
- 2) Save the Chinese sentences on these images as our first caption data of each collected image.
- 3) Perform a second-searching based on the keywords of the image captions we collected and similar image searching based on images we collected.
- 4) Extract the sentences from the second-searched images and save them as the caption data of the images we collected at the first time.

Using this process, we collected 3,000 meme images, corresponding to 30,000 sentences (each meme image had 10 sentences). Word counts of most of the sentences were between 5–10, as shown in Figure 3.(a). The images were divided into four categories according to the popularity of memes: television characters (film stars), internet celebrities (social media influencers), animals, and animation. The ratio of these four categories was 13:6:3:5, as shown in Figure 3.(b)

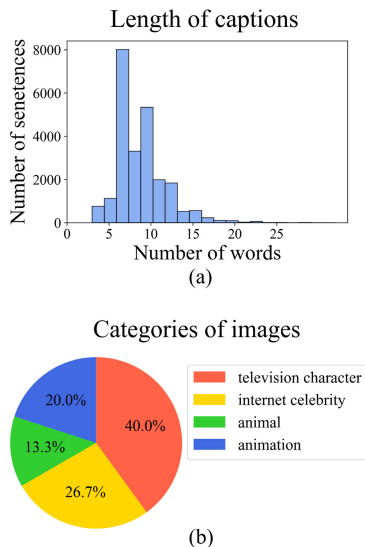


FIGURE 3. (a) Histogram of the number of words in each sentence in our Chinese meme dataset; (b) Pie chart of the categories of images in our Chinese meme dataset.

Image	Label	Captions
	59	<ul style="list-style-type: none"> <li>• 可爱又羞涩 ( Cute and shy )</li> <li>• 心花怒放 ( Burst with joy )</li> <li>• 一想到你就开心 ( Thinking of you means happiness )</li> <li>• ...</li> </ul>
	60	<ul style="list-style-type: none"> <li>• 我好紧张, 而且好害怕 ( So nervous and scared )</li> <li>• 瑟瑟发抖 ( Tremble with fear )</li> <li>• 你不要吓我, 我怕 ( Don't frighten me, I'm scared )</li> <li>• ...</li> </ul>
	61	<ul style="list-style-type: none"> <li>• 让我看看 ( Let me see )</li> <li>• 我都不敢看 ( Too scared to see )</li> <li>• 无法直视 ( I can't even look at this )</li> <li>• ...</li> </ul>
	62	<ul style="list-style-type: none"> <li>• 你懂我的意思的 ( You know what I mean )</li> <li>• 请开始你的表演 ( Start your show please )</li> <li>• 姐姐说笑了 ( You are overpraising me )</li> <li>• ...</li> </ul>
	63	<ul style="list-style-type: none"> <li>• 我太难了 ( So hard for me )</li> <li>• 我到底做错了什么 ( What have I done wrong )</li> <li>• 为什么要这么对我 ( Why treat me like this )</li> <li>• ...</li> </ul>
	64	<ul style="list-style-type: none"> <li>• 给你个白眼自己体会 ( You know why I roll my eyes )</li> <li>• 你在逗我? ( Are you kidding me? )</li> <li>• 小样儿, 你就继续装吧 ( Just keep acting bro )</li> <li>• ...</li> </ul>

FIGURE 4. Sample examples (images, label of images and part of captions of each image) from our Chinese meme dataset.

shows. Finally, to ensure one-to-one correspondence between the collected meme images and caption data, we referred to the label setting approach in “Dank Learning” [21]. From their experimental results, adding a specific label to the image of the training set does not have a positive impact on the meme generation. Therefore, we did not set other specific labels for the meme images of our dataset; instead, we used the serial numbers representing the images, which correspond to the same number allocated to the caption. An example of our dataset is shown in Figure 4.

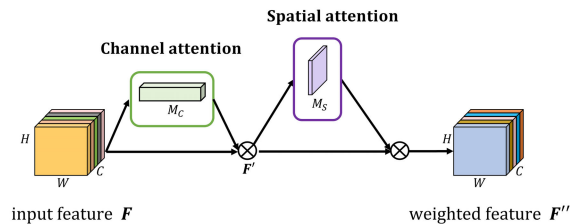


FIGURE 5. Cbam block added to our image extraction model. The input feature  $F$  represents the feature map of the first or last convolutional layer, Cbam extracts the feature maps from the two dimensions of channel and spatial, and then combines them with the original feature map to obtain the weighted feature  $F''$  as the input of the next convolutional layer. The  $F'$  is the weighted feature of channel dimension.

### C. MODEL ARCHITECTURE

As shown in Figure 2, our Chinese image meme generation system contains two components: an encoder and a decoder. The encoder is a CNN model for image feature extraction, and the decoder is a language model for generating Chinese text. The encoder-decoder architecture allows us to generate meme sentences based on a single input image.

#### 1) ENCODER

In this Chinese meme generation task, we used ResNet-50 [35] as our encoder to extract image features. Because the image classification result is not required, we removed the last fully connected layer of ResNet-50. We use the high-level semantic information of the images extracted from the last hidden layer of ResNet-50 as encoder output. To enhance the performance of the encoder for extracting high-level information, we adopted the convolutional block attention module (Cbam) [46], which is a lightweight module that can be integrated into CNN architectures smoothly. Moreover, Cbam can extract informative features by blending cross-channel and spatial features so that the model can learn “what” and “where” to attend in the channel and spatial dimensions, respectively, to refine the feature maps and improve performance. The input features go through channel attention first, followed by spatial attention. To use the pre-trained parameters on the ImageNet dataset [47], we did not modify the internal structure of ResNet-50, but added Cbam blocks to the first and last convolution layers.

When an intermediate feature map  $F_1 \in \mathbb{R}^{C \times H \times W}$  was input to the Cbam block, Cbam derives a 1D channel attention map  $M_c \in \mathbb{R}^{C \times 1 \times 1}$  and a 2D spatial attention map  $M_s \in \mathbb{R}^{C \times 1 \times 1}$ , as illustrated in Fig. 5. This process can be expressed as follows:

$$\begin{aligned}
 F' &= M_c(F) \otimes F \\
 F'' &= M_s(F') \otimes F'
 \end{aligned} \tag{1}$$

where  $\otimes$  denotes element-wise multiplication.  $F'$  is the result of channel attention, and  $F''$  is the final refined output. Thus, two attention calculations were performed on the first and last convolutional layers of the model. Relying on the attention

module, the model can learn the important parts of an image for the text generation process.

## 2) DECODER

In our Chinese meme generation task, the role of the decoder was to generate text based on image features as well as to generate humorous captions. Here, we chose the GPT-2 architecture [27] based on a multi-layer transformer decoder as our base Chinese text generative model because of its expression ability, as it has been certified in text generation tasks because of its self-attention capabilities [27], [40]. We used this resulting model to solve the problem of expressing humor in the caption to a certain extent. To avoid the influence of image noise information on language model learning, we referred to the experimental results of [48], where the image features did not participate in the word encoding process, but only affected the final word prediction. We used GPT-2, which contains  $H$  parallel heads, and each head  $h_i$  corresponds to an independent scaled dot-product attention function proposed in Transformer [40], which can be computed as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

where  $Q \in \mathbb{R}^{d_m \times d_k}$ ,  $K \in \mathbb{R}^{d_m \times d_k}$ , and  $V \in \mathbb{R}^{d_m \times d_v}$  are the query, key, and value matrices, respectively, and  $d_m$  is the dimension of the model.

Then, the attention results for each head  $h_i$  were composed by a linear transformation  $W^o$ , and the process was formulated as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(h_1, \dots, h_H)W^o \quad (3)$$

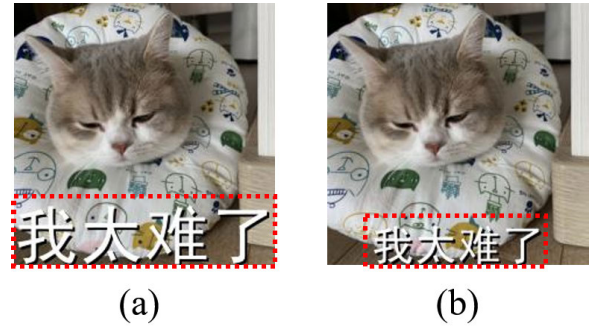
When the words were input to the decoder, we obtained the output features from the GPT-2 model and the input embedding; here, we added a multilayer perceptron (MLP) [49] with two hidden layers to merge image and word feature vectors into a single feature embedding in the default 768-dimension of the GPT-2 model. As a result, the image features can affect the prediction of the next word through a linear layer. Given a ground truth sequence  $y_{1:T}^*$ , the model is trained by minimizing the cross-entropy loss, illustrated as follows:

$$L_{XE}(\theta) = - \sum_{t=1}^T \log(p_{\theta}(y_t^* | y_{1:t-1}^*)) \quad (4)$$

where  $\theta$  is the parameter of the model, and  $y_t^*$  is the prediction.

## D. POST PROCESSING

After the meme captions are generated, to create the image meme following the form of that circulating on the Internet, we must combine the generated captions with the input images, which was our final post processing stage. By observing the styles of Chinese image memes that exist on the Internet, we found that most of the text is located at the bottom of the image and the text color is set to a bright color or white. In addition, the white text is usually placed on a black shadow to make it look distinct compared to the image. According to



**FIGURE 6. (a) Inappropriate example of text; it is a bit larger than the image. (b) Adjusted example. Set an appropriate font size and center the text.**

the observed characteristics, we generated and added captions to the images to obtain the final image memes following these three steps:

- 1) When inputting one image into the model, get the width  $I_{width}$  and the height  $I_{height}$  of the input image  $I$ . The number of text words  $t_n$  is calculated when the caption  $t$  is generated by the model prediction.
- 2) To put the generated caption on one line like general memes, calculate the font size  $t_f$  of the text.
- 3) Set the text color to white and add black shadows for each word.

For step (2) of calculating the text font size, first, we use the width of the image  $I_{width}$  divided by the number of words  $t_n$  to determine the font size of the text. Then, we can obtain the text position coordinates  $(t_x, t_y)$  through  $(0, I_{height} - t_f)$ .

In addition, we adjusted some special cases that were not suitable for this method. When the width of the input image was large with a small number of words, the text size seemed to occupy a large part of the image, as shown in Figure 6(a). For this example, we set the font size to an appropriate fixed size  $f$  and centered the text by calculating  $(I_{width} - f \times t_n) / 2$ ,  $(I_{height} - f)$ . An adjusted example is shown in Figure 6(b).









## IV. EXPERIMENTS

### A. DATASET

As mentioned in III-B, we present a dataset that contains 3,000 images, each image corresponding to 10 sentences, and a total of 30,000 image sentence pairs data were used for the experiments. We divided the dataset into training, validation, and test sets according to a ratio of 8:1:1.

### B. EXPERIMENTAL DETAILS

The encoder we used was the ResNe-50 pre-trained on the ImageNet dataset [47] for image feature extraction, and the decoder we used is the GPT-2 model, which contains 12 layers. The word embedding dimension was 768, and it had 12 attention heads to generate captions. In addition, we used the Chinese tokenizer presented by BERT [41] to add the start token '[CLS]' and the end token '[SEP]' for each text. The entire model was trained with cross-entropy loss for

Image	Generated captions	Image	Generated captions
	Without attention: 我太难了 (I'm too hard) With attention: 不敢相信自己的眼睛(Can't believe my eyes)		Without attention: 怪不好意思的 (So embarrassed) With attention: 我好想你啊 (I miss you so much)
	Without attention: 你说什么? (What did you say?) With attention: 你怎么回事小老弟(What's the matter with you, bro)		Without attention: 我不想跟你讲话 (Don't want to talk with you) With attention: 你这样是没有好下场的 (This is not going to end well for you)
	Without attention: 我都不知道怎么了 (I don't know what's wrong) With attention: 表情开始正经起来 (Expression turns decent)		Without attention: 不想理你 (Don't bother me) With attention: 你看我想理你吗 (Do you think I want to talk with you?)
	Without attention: 你不要搞事情啊 (Don't mess up) With attention: 看我嫌弃你的眼神 (Look my disgusted face)		Without attention: 这时候有点小忧伤 (Kind of sad now) With attention: 我没有不开心, 我只是不快乐 (I am not upset, I am just unhappy)

**FIGURE 7.** Captions generated by models without and with attention. It can be seen from the results that the generated captions using model with attention are more meaningful and contain rich information of corresponding image.

20 epochs using an initial learning rate of  $1e-5$ . Adam [50] was used as the optimizer. A beam search was performed, with beam size of 5. We trained our model on an NVIDIA GTX-1080 GPU.

### C. EVALUATION METRICS

We used the BLEU score [51], which is used to evaluate the difference between the model-generated captions and the actual sentences to evaluate the quality of the generated captions. The perspective of the good quality of a meme is subjective and varies among people. The focus of memes is to express images in a humorous way. To the best of our knowledge, there are no known automatic evaluation metrics for evaluating the quality of memes. A fairly reliable technique is human evaluation by a set of raters to evaluate the quality of a meme and provide a subjective score. We referred to the evaluation methods in Memeify [22] and performed two different user surveys for this study. We considered Dank learning [21] meme generator as a baseline model for comparative evaluation. This is because it is the only one that directly generates memes based on image input, which is similar to our model structure. Because the current dank learning model did not apply to Chinese meme generation results, we used our Chinese meme dataset to train the Dank Learning model to obtain the Chinese meme results, and then compared it with the results of our model. We also compared the memes that exist on the Internet. Additionally, Internet image memes have many formats and styles, such as font colors and text positions. Therefore, to eliminate the influence of other factors on the evaluation results, we separated the images and captions to evaluate them in the same format. We employed 20 users who used Chinese memes daily to conduct the two user surveys.

### D. RESULTS AND ANALYSIS

#### 1) CAPTION GENERATION RESULTS

We evaluated the results of the proposed model with and without attention to BLEU-1, BLEU-2, BLEU-3, and

**TABLE 1.** The results of the proposed model with and without attention on standard evaluation metrics: BLEU-1, BLEU-2, BLEU-3, and BLEU-4. The model with attention received better scores than those without attention.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4
without attention	55.2	28.6	13.8	3.2
with attention	64.4	34.6	21.4	9.8

BLEU-4 metrics in Table 1. The higher the score (close to 1), the more it shows that the generated caption is closer to the actual sentence. We find that the model with the Cbam attention layer has better BLEU scores than the base no attention model. According to our analysis and as mentioned in [46], by calculating not only the attention weights of the channel dimension, but also the spatial dimension to allow the model to learn a better feature map representation, so it generates better results than without attention. The results obtained from these two models are shown in Figure 7. We used the best performing model to generate memes for user surveys.

#### 2) RESULTS OF HUMAN EVALUATION STUDIES

##### a: USER RATING

To evaluate whether users were satisfied with and would use the generated memes, we conducted a scoring survey. We randomly selected 20 generated memes from our model, 20 memes on the Internet, and 20 generated memes from the Dank Learning model, without telling users whether they were real or generated. The users scored the memes based on whether the memes could express the content that the images were meant to express and how humorous they were. Scores ranged from 0 (completely non-compliant, i.e., not funny at all) to 9 (totally compliant, i.e., very funny). Finally, we averaged the scoring results, as listed in Table 2. We observed that the generated memes were almost comparable to the real memes, and in some cases, better, demonstrating that the quality of our generated memes was almost at the same level as that of the real memes on the Internet. Moreover, in response to the question “Do you want use the generated

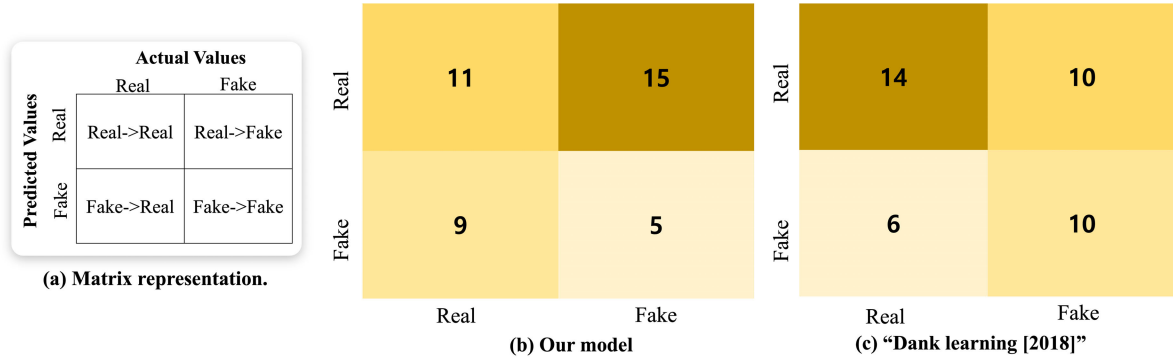


FIGURE 8. a) Explains what each element represents in our confusion matrices; (b) and (c) are confusion matrices for our model and the "Dank Learning [2018]" model.



- (a) : You are challenging me
- (b) : I think you should kiss me
- (c) : I'm so hungry
- (d) : I have never felt so embarrassed like this
- (e) : I am a little confused
- (f) : Is he out of mind?
- (g) : Your talk is so great
- (h) : I don't want to go to school
- (i) : Follow my lead
- (j) : No idea so just dance

FIGURE 9. Chinese image memes generated by our method.

TABLE 2. Human evaluation scores of real Internet memes and memes generated by our model. The actual score of the generated memes is slightly higher than that for the real memes; this does not mean that the generated memes are better than the real ones, but it can prove that they have a high probability of being liked by users.

Memes	Real Memes	Ours
Average score	6.54	6.83

memes by our model on the SNS," 80 percent of the users reported that they would like to use the memes generated by our model. Therefore, it could be concluded that most users were satisfied with the quality of the memes generated by our model.

*b: DISCRIMINATION ABILITY OF USERS*

We wanted to know whether the memes that we generated were considered to be real by users. In this survey, we divided the participants into two groups: the first group comprising 10 users were shown 20 memes generated by our model and 20 real memes from our dataset, and were asked to identify which ones were real and which were generated. In the second group, we showed another 10 users 20 generated memes from the Dank Learning model and 20 real memes from our dataset. We asked users to distinguish between the real and generated memes. Then, we generated statistics and conducted analysis of the results of the two groups, with reference to the evaluation method of Memeify [22], and drew



**TABLE 3. The evaluation results of baseline model and our model based on human evaluation scores.**

Methods	Accuracy	Precision	Recall	F1-score
Dank Learning[2018]	60	58.3	70	63.6
Ours	40	42.3	55	47.8

the confusion matrices for both the Dank Learning model and our model, as shown in Figure 8.

The confusion matrix shows that 75 percent of users classified our generated memes as real. This means that the memes generated by our model are largely regarded as real Internet memes. At the same time, based on the confusion matrix of the Dank Learning model, half of the memes were regarded as real memes, and the number of real memes judged to be fake was lower than that of memes generated by our model. This shows that, compared with real memes, the memes generated by our model are more confusing for users than those generated by the Dank learning model. This is because we rely on a more improved transformer-based architecture [27] with a self-attention mechanism and multiple-task adaptability to generate captions instead of the traditional RNN [35] architecture used in Dank Learning.

To further confirm the performance of our model, we used four metrics: accuracy, precision, recall, and F1-score, to analyze, based on this survey result, the results of these evaluation metrics as shown in Table 3.

It can be seen from the results that the memes generated by our model led to users being more likely to make mistakes in judgment. This demonstrates that the generated memes are sufficiently humorous to make it difficult for users to distinguish between real and generated memes.

As shown in Figure 9, we select some examples from the generated Chinese image memes using our model. We can see that our model can generate readable text content and maintain rich semantic information about the images. The generated caption can successfully describe what the image is intended to express.

## V. CONCLUSION

In this paper, we explain our Chinese meme generation model, which can creatively generate humorous Chinese captions given any image by employing ResNet-50 for image feature extraction, combined with GPT-2 for Chinese meme text generation. To generate image memes, we visualized the generated captions on the images in the general style of Internet image memes. To capture humorous features, we collected a dataset consisting of 30,000 Chinese texts with 3,000 Chinese meme images. Further, we provided an in-depth qualitative analysis based on user studies and proved that the Chinese memes generated by our model cannot be differentiated from real ones.

However, when evaluating the degree of humor of our generated Chinese memes, the user surveys made it clear that sense of humor varies greatly from person to person. A direct metric for evaluating the degree of humor is neces-

sary. However, the existing suitable metric is rare. In future work, we will study other suitable metrics to evaluate the results of our model. For example, the metrics for sentiment image and text analysis are used to determine whether the emotions expressed by the image and text are consistent. In addition, our current system can only generate memes with text at the bottom of the image. We would like to extend our system to learn the appropriate position of the text to generate diversified memes.

## REFERENCES

- [1] *Definition of Meme*. Accessed: Sep. 30, 2021. [Online]. Available: <https://www.merriam-webster.com/dictionary/meme>
- [2] (2006). *Twitter*. [Online]. Available: <https://twitter.com>
- [3] (2004). *Facebook*. [Online]. Available: <https://facebook.com/>
- [4] (2009). *Sina Weibo*. [Online]. Available: <https://www.weibo.com/>
- [5] (2011). *Wechat*. [Online]. Available: <https://www.wechat.com/>
- [6] A. Voulozimos, N. Doulamis, and A. Doulamis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–13, Jul. 2018.
- [7] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," *J. Amer. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 544–551, 2011.
- [8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
- [9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25, Dec. 2012, pp. 1097–1105.
- [10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
- [11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.
- [12] S. Edunov, M. Ott, M. Ranzato, and M. Auli, "On the evaluation of machine translation systems trained with back-translation," 2019, *arXiv:1908.05204*.
- [13] Y. Jia, R. J. Weiss, F. Biadsy, and W. Macherey, "Direct speech-to-speech translation with a sequence-to-sequence model," in *Proc. Interspeech*, 2019, pp. 1–5, doi: [10.21437/interspeech.2019-1951](https://doi.org/10.21437/interspeech.2019-1951).
- [14] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen, "Enhanced LSTM for natural language inference," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 1657–1668.
- [15] A. Wei Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le, "QANet: Combining local convolution with global self-attention for reading comprehension," 2018, *arXiv:1804.09541*.
- [16] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.
- [17] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.
- [18] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4894–4902.
- [19] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4613–4621.
- [20] H. Li, P. Wang, C. Shen, and A. V. D. Hengel, "Visual question answering as reading comprehension," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6319–6328.
- [21] A. L. Peirson V and E. Meltem Tolunay, "Dank learning: Generating memes using deep neural networks," 2018, *arXiv:1806.04510*.
- [22] S. R. Vyalla and V. Udandarao, "Memeify: A large-scale meme generation system," in *Proc. 7th ACM IKDD*, 2020, pp. 307–311.
- [23] (2018). *Dank Learning App*. [Online]. Available: <https://danklearning.com/>
- [24] (2018). *Memegenerator*. [Online]. Available: <https://www.memegenerator.net/>

- [25] K. Xu, "Show, attend and tell: Neural image caption generation with visual attention," *Comput. Sci.*, vol. 2015, pp. 2048–2057, Feb. 2015.
- [26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
- [27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
- [28] A. Sadasivam, K. Gunasekar, H. Davulcu, and Y. Yang, "MemeBot: Towards automatic image meme generation," 2020, *arXiv:2004.14571*.
- [29] W. Y. Wang and M. Wen, "I can has cheezburger? A nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2015, pp. 355–365.
- [30] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.
- [31] Y. Chen, Z. Wang, B. Wu, M. Li, H. Zhang, L. Ma, F. Liu, Q. Feng, and B. Wang, "MemeFaceGenerator: Adversarial synthesis of Chinese meme-face from natural sentences," 2019, *arXiv:1908.05138*.
- [32] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014, *arXiv:1406.2661*.
- [33] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 27, 2014, pp. 3104–3112.
- [34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
- [35] A. Graves, "Long short-term memory," in *Supervised Sequence Labeling With Recurrent Neural Networking*. Berlin, Germany: Springer, 2012, pp. 37–45.
- [36] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
- [37] L. Huang, W. Wang, J. Chen, and X. Y. Wei, "Attention on attention for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Dec. 2019, pp. 4634–4643.
- [38] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10971–10980.
- [39] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
- [40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
- [41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
- [42] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," Tech. Rep., 2018. [Online]. Available: [http://openai-assets.s3.amazonaws.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](http://openai-assets.s3.amazonaws.com/research-covers/language-unsupervised/language_understanding_paper.pdf)
- [43] (2018). *Reddit Dank Memorial Dataset*. [Online]. Available: <https://www.kaggle.com/sayangoswami/reddit-memes-dataset>
- [44] (2018). *Meme Generator Data Set*. [Online]. Available: <https://www.kaggle.com/electron0zero/memegenerator-dataset/home>
- [45] (2016). *52 Douyu*. [Online]. Available: <https://www.52douyu.cn>
- [46] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
- [47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
- [48] M. Tanti, A. Gatt, and K. P. Camilleri, "Where to put the image in an image caption generator," *Natural Lang. Eng.*, vol. 24, no. 3, pp. 467–489, May 2018.
- [49] M. W. Gardner and S. R. Dorling, "Artificial neural networks (the multi-layer perceptron)—A review of applications in the atmospheric sciences," *Atmos. Environ.*, vol. 32, nos. 14–15, pp. 2627–2636, Aug. 1998.
- [50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
- [51] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Conf. Assoc. Comput. Linguist. Meeting*, 2002, pp. 311–318.



**LIN WANG** received the B.S. degree from the Department of Computer Science and Technology, Northeastern University, China, in 2017, and the M.S. degree from the Department of Computer Science and Engineering, Korea University, in 2020, where she is currently pursuing the Ph.D. degree. Her current research interests include deep learning, image processing, and computer vision.



**QIMENG ZHANG** received the B.S. degree from the Department of Computer Science and Technology, Zhengzhou University, in 2015, and the M.S. degree from the Interdisciplinary Program in Visual Information Processing, Korea University, in 2018, where she is currently pursuing the Ph.D. degree. Her current research interests include geometry processing, physically-based simulation, and virtual reality.



**YOUNGBIN KIM** (Member, IEEE) received the B.S. and M.S. degrees in computer science and the Ph.D. degree in visual information processing from Korea University, in 2010, 2012, and 2017, respectively. From August 2017 to February 2018, he worked as a Principal Research Engineer with Linwalks. He is currently an Assistant Professor with the Graduate School of Advanced Imaging Science, Multimedia and Film, Chung-Ang University. His current research interests include data science and deep learning.



**RUIZHENG WU** received the B.E. degree from the Department of Computer Science and Engineering, South China University of Technology, in 2017, and the Ph.D. degree from the Computer Science and Engineering Department, The Chinese University of Hong Kong, in 2021. His current research interests include video instance/object segmentation, generative adversarial networks, and deep learning networks.



**HONGYU JIN** received the B.S. degree from the Department of Computer Science and Engineering, Korea University, in 2021, where he is currently pursuing the master's degree with the Interdisciplinary Program in Visual Information Processing. His current research interests include deep learning-based computer vision and augmented reality.



**HAOKE DENG** received the B.S. degree from the College of Software, Harbin Engineering University, in 2020. He is currently pursuing the master's degree with the Department of Computer Science and Engineering (Software), Korea University. His current research interests include physical-based AR interaction and deep learning-based computer vision.



**PENGCHU LUO** received the B.S. degree from the Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, in 2020. She is currently pursuing the M.S. degree in computer science and engineering with Korea University. Her current research interests include geometry processing, physically-based simulation, and deep learning.



**CHANG-HUN KIM** received the B.A. degree in economics from Korea University, in 1979, and the Ph.D. degree from the Department of Electronics and Information Science, Tsukuba University, Japan, in 1993. After graduation, he joined the Korea Advanced Institute of Science and Technology as a Research Scientist, where he was involved in many national research projects in the area of computer aided design and geometric modeling. He is currently a Professor with the Department of Computer Science and Engineering, Korea University. His current research interests include fluid animation and mesh processing. He is a member of the IEEE Computer Society and ACM.

...