

Knowledge-Guided Cross-Topic Visual Question Generation

Hongfei Liu^{1,2}, Guohua Wang^{1,2}, Jiayuan Xie³, Jiali Chen^{1,2}, Wenhao Fang^{1,2},
Yi Cai^{1,2,*}

¹School of Software Engineering, South China University of Technology

²Key Laboratory of Big Data and Intelligent Robot

(South China University of Technology) Ministry of Education

³Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China

{202221045918, segarychen, sewenhaofang}@mail.scut.edu.cn

{ghwang, ycai}@scut.edu.cn, jiayuan.xie@polyu.edu.hk

Abstract

Visual question generation (VQG) task aims to generate high-quality questions based on the input image. Current methods primarily focus on generating questions containing specified content utilizing answers or question types as constraints. However, these constraints make it challenging to control the topic of generated questions (e.g., conversation or test subject topics) for various applications. Thus, it is necessary to utilize topics as constraints to guide question generation. Considering that there are many topics and it is almost impossible for human annotations to cover them, we propose the cross-topic learning VQG (CTL-VQG) task, which aims to generate questions related to unseen topics in cross-topic scenarios. In this paper, we propose a knowledge-guided cross-topic visual question generation (KC-VQG) model to extract unseen topic-related information for question generation. Specifically, an image-topic feature extractor is introduced in our model to extract topic-related intuitive visual features; an image-topic knowledge extractor is used to extract and select the most appropriate topic-related implicit knowledge from large language models for generating questions. Extensive experiments show that our model outperforms baselines and can effectively generate unseen topic-related questions in cross-topic scenarios.

Keywords: Cross-Topic, Visual Question Generation, Knowledge-Guided

1. Introduction

Visual question generation (VQG) task endeavors to generate questions based on the given image. VQG has obtained significant attention in both the computer vision and the natural language processing areas due to its various potential applications in intelligent education systems (He et al., 2017) and dialogue systems (Mostafazadeh et al., 2016), etc. For these various applications, the generated questions usually need to contain specified content based on certain constraints, e.g., test points in education and topics in conversations. To this end, existing VQG methods employ answers (Krishna et al., 2019; Xie et al., 2021; Xu et al., 2021) or question types (e.g., “what”, etc) (Fan et al., 2018) as constraints to guide question generation. Although subject to some constraints, it is also hard to generate specific questions that are suitable for various applications. As shown in Figure 1, given an answer (i.e., white) as constraint may generate dialogue that is inappropriate for the contextual topic, or generate a question that is irrelevant to the exam topic “Geography” based on the question type “Is”. Due to the lack of constraints on topics, it is difficult for these VQG methods to generate appropriate questions. Thus, it is necessary

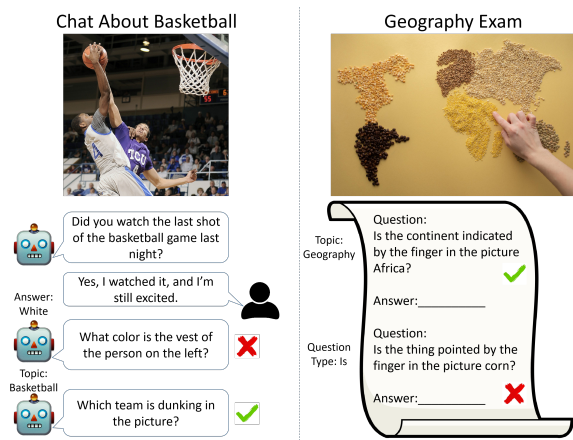


Figure 1: A comparison between answer constraints and topic constraints in VQG task.

to utilize topics as constraints to guide question generation (Duan et al., 2008). However, there are many topics in question and it is almost impossible to fully cover them with human annotations. In this paper, we propose the cross-topic learning VQG (CTL-VQG) task. The task aims to generate questions across topic scenarios, i.e., the ability to generate questions that are learned within a limited number of labeled topics and effectively extends to those unlabeled topics, especially situations that

* Corresponding author

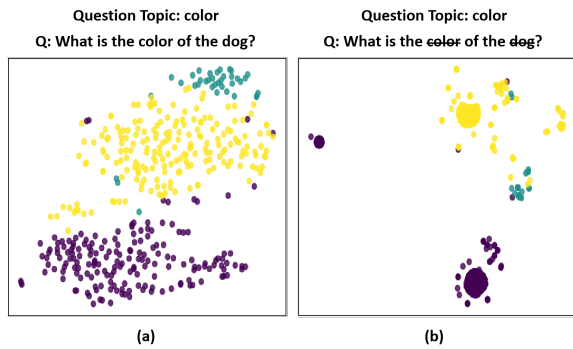


Figure 2: Question embedding distribution in different question types with the question topic “color”. Each point represents the coordinates of a question embedding after dimensionality reduction and normalization. The points are color-coded to represent questions of different question types. Figure (a) represents the original question embedding. Figure (b) represents the embedding of the question without topic-related content.

are uncommon or require expertise to label.

To solve this CTL-VQG task, a VQG model should learn some valid paradigms from the data of the annotated topics. Following the research of (Duan et al., 2008), humans typically regard each question as being divisible into two components, i.e., structural information and content information. Specifically, structural information is affected by question type such as “which” or “is”, while content information is mainly related to the topic. As shown in Figure 2(a), the distribution of all questions under the same topic “color” is shown, and the distribution of each question type is evenly distributed. When we delete specific content such as the object “dog” described by the topic “color”, as shown in Figure 2(b), the questions with each question type are aggregated together separately. This indicates that cross-topic learning mainly captures content related to a new topic without annotation while maintaining structural information.

To this end, we propose a knowledge-guided cross-topic visual question generation (KC-VQG) model, which aims to generate a question related to an unseen topic with specified types under cross-topic conditions. The KC-VQG model consists of three components, i.e., image-topic feature extractor (IT-FE), image-topic knowledge extractor (IT-KE), and topic and type-guided question decoder (TT-GQD). For an unseen topic, we consider that topic-related content can be divided into intuitive visual features and implicit knowledge, e.g., an object “football” in topic “sport” based on intuitive visual features, and an object “lamp” in topic “function” based on a knowledge “lamp can be used in dark rooms”. To capture intuitive visual features, the ITFE employs a topic-aware attention mechanism

to capture the visual regions related to the topic in the image. Then, we leverage ITKE to extract topic-related implicit knowledge for question generation from large language models (LLMs) e.g., ChatGPT (OpenAI, 2023). Considering that not all LLMs’ generated knowledge is suitable for capturing topic-related content, the ITKE subsequently introduced a knowledge discriminator to evaluate the suitability of each piece of generated knowledge for formulating the question. Finally, the TT-GQD employs a distilled GPT-2 model to generate questions. For the structural information, as the question type is a trainable component, we train the model by incorporating the question type into the prompt. For the content information, we first add extracted knowledge to the prompt to allow the model to obtain topic-related implicit knowledge. Additionally, we apply the features obtained by the topic attention mechanism to GPT-2 to promote the generation of topic-related intuitive visual features.

In summary, our contributions are as follows:

- To the best of our knowledge, our work is the first one to explore cross-topic learning in the VQG (CTL-VQG) task, solving the problem that existing methods fail to generate appropriate questions for unseen topics. The task aims to generate topic-related questions even in the absence of corresponding topic annotations, which can be used in various applications.
- To address the challenge of the CTL-VQG that a VQG model fails to obtain topic-related content through training in cross-topic learning, we propose to use the topic-related visual features and additional knowledge from LLMs. Furthermore, we introduce a knowledge discriminator, designed to filter out topic-specific knowledge that is suitable for generating questions.
- Experiments show that existing VQG models perform poorly in cross-topic conditions. Meanwhile, our proposed model achieves state-of-the-art results in most of the topics as well as in the overall evaluation.

2. Related Work

2.1. Visual Question Generation

Visual question generation (VQG) can be divided into two categories: unconstrained VQG and constrained VQG. Unconstrained VQG generates questions based solely on the input image (Mostafazadeh et al., 2016; Bi et al., 2022), without any additional constraints. On the other hand, constrained VQG incorporates constraints such as answers (Krishna et al., 2019; Xie et al., 2021; Xu et al., 2021; Chen et al., 2021) to control the model

and generate more specific and accurate questions. Specifically, Krishna et al. (2019) propose to use the answer as a constraint to generate questions related to the image and the given answer. Xie et al. (2021) utilizes a graph network to identify all key objects related to the answer. Xu et al. (2021) propose to use the innovative radial graph convolutional networks to quickly identify the central regions in images related to the answer. Bi et al. (2022) propose a method to generate questions with controllable difficulty by building a core scene graph and using path searching with multi-hop inference. Xie et al. (2022) introduced external knowledge to generate knowledge-related visual questions. Fan et al. (2018) propose using question types as a guide to generate diverse questions. Vedde et al. (2022) propose using answer categories as constraints to generate questions. Chen et al. (2023) introduce a causal perspective and leverage external knowledge to mitigate spurious correlations, demonstrating superior performance over existing methods on VQA v2.0 and OKVQA datasets.

2.2. Cross-Domain Learning

Cross-domain learning refers to a problem setting in machine learning where the model is evaluated on generating content for a domain that was not included in the training dataset (Chao et al., 2018). This approach aims to address the issue of limited data availability by enabling models to learn from previously unseen domains, receiving significant attention from researchers in recent years (Zhang et al., 2021). Over the past few years, numerous advances have been proposed in the field of cross-domain learning (Chen et al., 2022; Ding et al., 2022; Huang et al., 2022; Jain et al., 2022; Gera et al., 2022; Fang et al., 2022). Zheng et al. (2021) propose the cross-domain instance segmentation task, which can perform instance segmentation in fields like medical imaging without field-related training data. Esmaeilpour et al. (2022) utilize the pre-trained CLIP model to solve the cross-domain out-of-distribution detection problem. Tewel et al. (2022) propose combining the visual-semantic model with a large language model to generate image captions in the cross-domain case. Tian et al. (2022) propose a Transformer-based approach to solve the cross-domain sketch-based image retrieval task.

To the best of our knowledge, we are the first to explore cross-topic scenarios for question generation in the VQG task.

3. Methodology

In this CTL-VQG task, we first classify the questions into N different topics $T = \{t_1, t_2, \dots, t_N\}$, e.g.,

sports, color, etc. For a given image I , answer A , question type $w_j \in \{w_1, w_2, \dots, w_M\}$ (e.g. what, when, how, etc.), and a question topic $t_i \in T$, our objective is to enable the model to generate questions specific to the unseen topic t_i , while using the other topics $T' = \{t_k | k \neq i\}$ data for training. Our model mainly focuses on capturing content related to the unseen topic t_i in I , while maintaining structural information corresponding to w_j . Specifically, the overall framework of our model is shown in Fig. 3, which consists of three components: (i) image-topic feature extractor, which is used to extract topic-related intuitive visual features for a given image; (ii) image-topic knowledge extractor, which is utilized for the extraction of topic-related implicit knowledge from large language models (LLMs); (iii) topic and type-guided question decoder, which is used to generate questions based on topic-related information (i.e., topic-related intuitive visual features and topic-related implicit knowledge) and question type.

3.1. Image-Topic Feature Extractor

The IT-FE is employed to identify topic-related intuitive visual content that exhibits semantic relevance to the topics portrayed in the image. For instance, for a question with sport topic such as “What sport is the man doing?”, the model needs to focus on sport-related visual regions (e.g., basketball) in the image. The module comprises three components: the visual encoder, the text encoder, and the attention module. Specifically, the visual encoder extracts image features, the text encoder extracts textual topic features, and the attention module enables the model to prioritize topic-related intuitive visual regions.

Visual Encoder: The visual encoder is utilized to extract the features of the image. To acquire visual features that align semantically with the topic features, we utilize a pre-trained CLIP ViT-B/16 visual encoder to encode the image. Specifically, the image is resized to 224×224 first and divided into $P = 14 \times 14 = 196$ patches with the size of 16×16 . Each patch obtains a visual feature v_p by the CLIP visual encoder. Therefore, the image features can be denoted as $V = \{v_p\}_{p=1}^{196}$.

Text Encoder: The text encoder is used to extract the text features of the input topic. To acquire word-level features of the question topic that align semantically with the visual features, we employ the pre-trained CLIP text encoder. This model, characterized by a multi-level transformer structure, encodes the word embeddings of the question topic t_i into a 768-dimensional vector e_i . To ensure that the extracted visual features and text features of the topic are in the same vector space, the pre-trained CLIP visual encoder and text encoder are kept frozen during training phases.

Attention Module: The attention module operates

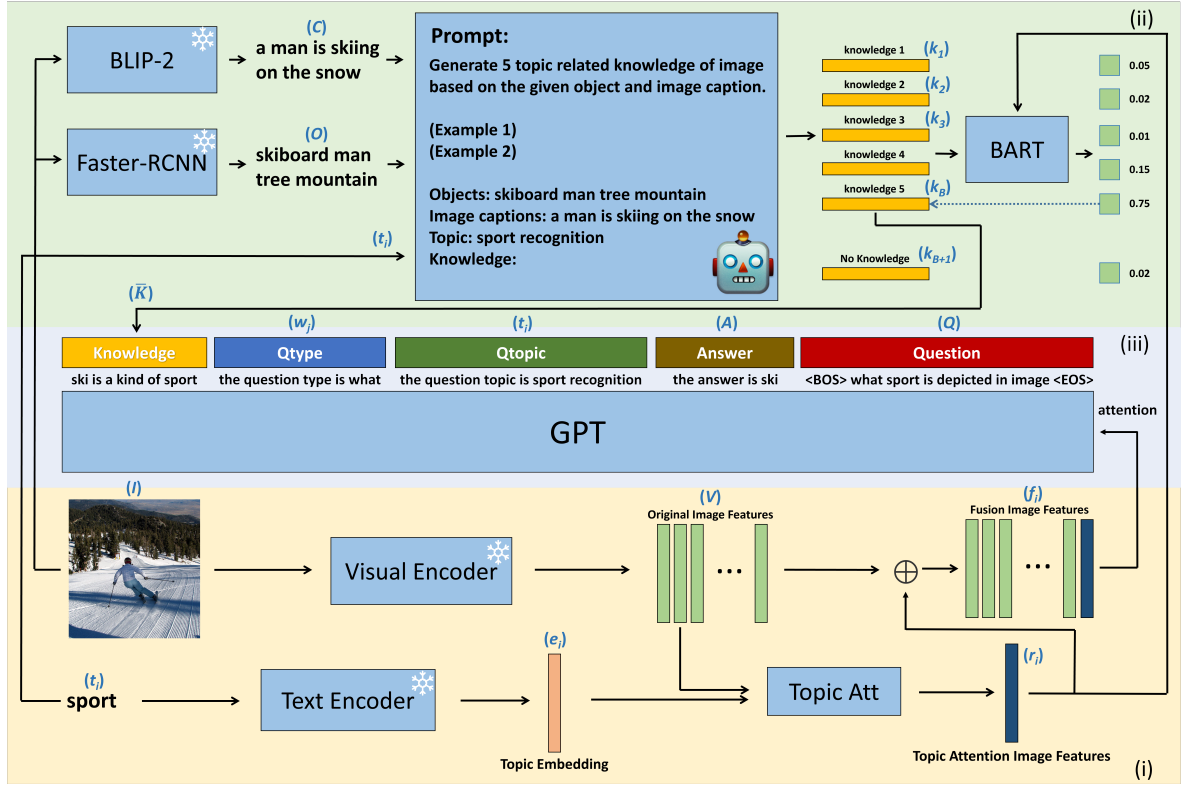


Figure 3: Overview of our KC-VQG model. It contains three components: (i) image-topic feature extractor, (ii) image-topic knowledge extractor, (iii) topic and type-guided question decoder.

by focusing on the extracted visual features and topic features to derive topic-related visual features. Given that these features are aligned within the same vector space, the attention module can capture image regions related to the topic even in cross-topic scenarios. Specifically, we employ the question topic embedding e_i , to weigh the visual features V , resulting in topic-related visual feature r_i . We splice e_i and v_p and input them to the multi-layer perceptron (MLP) model to compute the weights of the p -th visual feature w_p^i :

$$w_p^i = MLP(e_i \oplus v_p), \quad (1)$$

where \oplus is the concatenation operation. Next, we use the attention scores $w^i = (w_1^i, w_2^i, \dots, w_P^i)$ as weights to perform a weighted summation on the visual features $V = (v_1, v_2, \dots, v_P)$, to obtain the topic-related visual feature r_i :

$$r_i = \sum_{p=1}^P w_p^i v_p, \quad (2)$$

where v_p denotes the p -th visual feature.

To acquire the visual fusion feature related to the topic, we concatenate the topic-related visual feature r_i with the image features $V = (v_1, v_2, \dots, v_P)$ to obtain the fusion image feature, denoted as f_i :

$$f_i = V \oplus r_i. \quad (3)$$

3.2. Image-Topic Knowledge Extractor

Given that certain topic-related information cannot be directly obtained from the intuitive visual information in images, such as the relationship between the object “lamp” and the topic “function”, the IT-KE Module is employed to uncover implicit semantic knowledge related to topics. The module comprises two components: the knowledge generator and the knowledge discriminator. To be precise, the knowledge generator is employed to generate implicit knowledge pertinent to the topic, resulting in multiple candidate knowledge entries. Subsequently, the knowledge discriminator assesses and scores each knowledge candidate, facilitating the selection of the most appropriate knowledge for generating questions.

Knowledge Generator: The GPT-3.5 model exclusively accommodates textual input, so it becomes necessary to transform image input into a textual format. To this end, we employ the pre-trained BLIP-2 model to generate the image caption C , and the pre-trained faster-rcnn model to extract the 36 most prominent objects O in the image. We encode the image caption C , the objects O , and the topic t_i into prompts $[C, O, t_i]$ as input to GPT-3.5 to generate B different image-related topic knowledge candidates, denoted as $K = \{k_1, k_2, \dots, k_B\}$, for each image-topic pair. The detail of the specific

prompt content is shown in Fig. 3.

Knowledge Discriminator: Given the presence of redundant knowledge within the knowledge extracted from GPT-3.5, we introduce a knowledge discriminator to select the most appropriate knowledge for generating questions. Recognizing that certain questions do not require the utilization of knowledge, i.e., "What color is the woman's shirt?", we include a $B + 1$ knowledge option labeled as "no knowledge," denoted as k_{B+1} . Both the knowledge extractor outputs K and the "no knowledge" option k_{B+1} are considered within the knowledge discrimination process, denoted as $K' = \{k_1, k_2, \dots, k_{B+1}\}$. The knowledge discriminator evaluates and assigns scores to each knowledge option, ultimately selecting the most suitable knowledge, denoted as \bar{K} , for generating questions.

Specifically, we employ a bidirectional and autoregressive transformer (BART) (Lewis et al., 2020) model to encode the t -th knowledge k_t , utilizing the last hidden state as the knowledge feature g_t . We concatenate the knowledge feature g_t with the topic-related visual feature r_i obtained from the IT-FE module. Subsequently, we input this concatenated feature into an MLP to calculate the score of the t -th knowledge, denoted as s_t :

$$s_t = MLP(g_t \oplus r_i). \quad (4)$$

During the testing phase, we choose the knowledge option with the highest score \bar{K} to serve as the input knowledge for the question decoder.

3.3. Topic and Type-Guided Question Decoder

To leverage the exceptional capabilities of GPT (Brown et al., 2020), we utilize it as the question decoder. Specifically, we utilize a distilled GPT-2 model, which was pre-trained on a large-scale corpus of image-caption data, as our question decoder (Sammani et al., 2022). The question decoder takes four inputs: knowledge, question type, question topic, and answer. Knowledge helps the model generate questions related to the topic, while question type aids in understanding the question's structure. The input answer and question topic are employed to address the requirement for generating context-specific questions.

Specifically, we create the input sequence by incorporating information about the knowledge, question type, topic, and answer, in the format "the knowledge is K the question type is w_j the question topic is t_i the answer is A ". Then, we generate the question Q in an autoregressive manner, using the beginning-of-sequence token $\langle BOS \rangle$, followed by the question content, and ending with the end-of-sequence token $\langle EOS \rangle$. Since the generated questions aim to be related to the fused image features, we utilize the hidden state h_i at each time

step in the GPT model as a query and the image features f_i as the key and value to calculate the vanilla attention (Vaswani et al., 2017).

The model is trained using the cross-entropy objective to generate a sequence $y = y_1, y_2, \dots, y_T$ of T words as the question. The objective is to minimize the negative log-likelihood:

$$L = - \sum_{\theta=1}^T \log p(y_\theta | y_{<\theta}), \quad (5)$$

where $y_{<\theta}$ denotes the words before the θ -th word.

To acquire the necessary labels for knowledge discriminator training, during the training phase, we provide each knowledge option, namely, $K = \{k_1, k_2, \dots, k_{B+1}\}$, as input to the question decoder to generate questions. We then calculate the loss by comparing the generated questions with the ground truth, resulting in individual loss values, denoted as $L = \{l_1, l_2, \dots, l_{B+1}\}$. We assign a positive label to the knowledge corresponding to the smallest loss and assign negative labels to the other knowledge options. When updating parameters in the training phase, we specifically select the knowledge with the smallest loss for back propagation.

4. Experiment

4.1. Dataset

We conduct experiments on the TDIUC dataset (Kafle and Kanan, 2017). Each image in the dataset is accompanied by a collection of questions, each with corresponding topics and answer tags. We preprocess the dataset by eliminating duplicated questions. Additionally, we removed questions categorized as "absurd", indicating that they cannot be answered based on the content in the image. We conduct experiments using the dataset consisting of 11 topics and 15,814 questions. In each cross-topic scenario, we allocated all the data from the target topic to the test set, while including data from other topics in the training set. Subsequently, we further divided the training set into a training subset and a validation subset, maintaining an 8:2 ratio.

4.2. Evaluation Metrics

Automatic Evaluation Metrics. To evaluate the effectiveness of our model, we employed several widely used (Krishna et al., 2019; Xu et al., 2021) evaluation metrics, including BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015), and BERT-score (Zhang et al., 2020) to assess the quality of the generated questions.

Specifically, BLEU-4 and ROUGE-L metrics primarily focus on word-level coverage between generated questions and references. In contrast, the

Meteor metric considers the accuracy and recall based on the entire corpus, sentence fluency, and synonymy, making it more comprehensive than the BLEU metric. Additionally, the CIDEr metric assesses whether generated questions contain key information and has shown a stronger correlation with human assessments. The BertScore metric is used to evaluate the semantic similarity between the generated questions and references.

Human Evaluation Criteria. Automatic evaluation metrics have a limitation in that they can only be compared to ground truth questions. However, considering the diversity of reasonable questions, the comparison results with ground truth questions may not fully represent the quality of generated questions. Therefore, we utilize human evaluation as an additional assessment to overcome this limitation.

The following criteria are used for human evaluation: **Fluency (Flu)** is used to evaluate the coherence, fluency, syntax, and grammar of the generated questions. **Image relevance (Img_rl)** is used to evaluate whether the generated questions are related to the image content. **Answer relevance (Ans_rl)** is used to evaluate whether the generated questions can be answered based on the image content and whether they match the input answers. **Topic relevance (Top_rl)** is used to evaluate whether the generated questions are related to the input topic. Each criterion will be assessed on a scale ranging from 0 to 5, with higher scores indicating a closer alignment of the generated questions with the criteria.

4.3. Baselines and Ablation Models

To evaluate the effectiveness of our framework, we conducted a comparative experiment against the following baseline models, which fall into four categories: seq2seq-based, GCN-based, transformer-based, and knowledge-based models:

IMVQG (Krishna et al., 2019) utilizes CNN and LSTM models to generate a wider range of goal-oriented questions by maximizing the mutual information between the input answers and the generated questions.

VQG-GCN (Xu et al., 2021) apply GCN model for the VQG task. It is an answer-centric approach, which effectively models the associations between the answer and its relevant image regions.

MOAG (Xie et al., 2021) enhances question generation by incorporating a GCN to capture relationships between essential objects and unrelated objects within the image. This model enables the generation of questions that incorporate more visual context.

ClipCap (Mokady et al., 2021) is an advanced image captioning model. In this experiment, we trained it using questions and evaluated its output as a baseline. ClipCap uses CLIP as its visual

encoder, the GPT-2 model as its text decoder, and a transformer-based mapping network to map text-to-image vectors into text space.

KB-VQG (Xie et al., 2022) is a knowledge-based model that generates questions containing knowledge by retrieving knowledge in ConceptNet (Speer et al., 2017).

GPT-3.5 (OpenAI, 2023) uses the input directly from the knowledge extractor, including image caption, objects, and topic, we employ the GPT-3.5-turbo model to generate questions directly.

The existing models have not considered generating questions specifically tailored for cross-topic scenarios, leading to a limitation in producing high-quality questions in such contexts.

To ensure the fairness of the experiment, we concatenated the question type, question topic, and answer to form the input of the baseline model. This was done to ensure that the content of each model input was consistent across the experiments.

Moreover, to evaluate the effectiveness of our proposed module, we conducted the following ablation experiments:

KC-VQG w/o KD: The KC-VQG model generates questions without utilizing the knowledge discriminator to determine the most suitable knowledge for question formulation. During the testing phase, random knowledge is chosen as input for the decoder to generate questions.

KC-VQG w/o TA: The KC-VQG model generates questions by excluding topic-related visual feature r_i from the fusion image features f_i . The original image features V are inputted into the GPT model to calculate the vanilla attention.

4.4. Experiment Details

We implement our model by using the pytorch framework and train the model with a single GTX2080 Ti GPU. We utilize a CLIP model with a ViT-B/16 Transformer architecture, which was pre-trained on publicly available image-caption data, as our visual encoder (Radford et al., 2021). The distilled GPT-2 model we utilize was pre-trained on a large-scale corpus of image-caption pairs (Sammani et al., 2022). The hyper-parameters of our model are described as follows. We set the number of knowledge generated by knowledge generator B to 5. We set the hidden size of the 2-layer MLP in the text encoder to 512 and the hidden size of the 2-layer MLP in the attention module to 256. We train the model for up to 5 epochs using an Adamax optimizer (Kingma and Ba, 2015). We set the batch size to 128 and the learning rate to 2×10^{-5} .

Topic	Model	BLEU-3	BLEU-4	METEOR	ROUGE _L	CIDEr	BertScore
activity recognition	IMVQG (Krishna et al., 2019)	16.55	6.72	18.06	43.76	28.54	86.00
	VQG-GCN (Xu et al., 2021)	1.25	0.00	11.12	27.13	5.10	85.79
	MOAG (Xie et al., 2021)	6.70	0.00	14.13	51.08	4.85	86.34
	ClipCap (Mokady et al., 2021)	8.27	4.99	15.98	32.30	16.41	85.87
	KB-VQG (Xie et al., 2022)	15.63	6.43	16.75	41.74	41.25	86.76
	GPT-3.5 (OpenAI, 2023)	7.02	5.47	10.32	32.16	86.70	91.35
	KC-VQG w/o KD	12.36	9.36	12.37	39.27	53.07	88.51
	KC-VQG w/o TA	17.79	11.40	15.57	48.50	34.41	88.72
	KC-VQG(Ours)	27.65	17.24	20.90	53.93	50.45	92.33
attribute	IMVQG (Krishna et al., 2019)	8.11	2.41	11.78	34.80	1.96	85.15
	VQG-GCN (Xu et al., 2021)	1.60	0.00	11.86	28.90	2.35	87.81
	MOAG (Xie et al., 2021)	17.66	4.39	11.94	47.67	3.63	82.39
	ClipCap (Mokady et al., 2021)	9.21	3.58	13.35	32.25	9.56	84.48
	KB-VQG (Xie et al., 2022)	10.04	0.00	12.05	33.67	7.82	85.77
	GPT-3.5 (OpenAI, 2023)	5.74	2.26	9.57	30.80	21.49	87.78
	KC-VQG w/o KD	18.49	11.53	15.84	48.30	57.39	88.41
	KC-VQG w/o TA	18.19	11.24	15.20	50.99	48.03	88.99
	KC-VQG(Ours)	26.92	16.63	17.19	51.82	72.44	89.89
color	IMVQG (Krishna et al., 2019)	10.61	0.00	12.48	47.06	6.23	86.87
	VQG-GCN (Xu et al., 2021)	2.88	0.92	11.16	28.73	4.82	87.84
	MOAG (Xie et al., 2021)	44.15	36.19	26.56	62.55	6.51	88.57
	ClipCap (Mokady et al., 2021)	13.80	9.74	20.43	38.43	10.06	85.76
	KB-VQG (Xie et al., 2022)	9.99	0.00	12.86	46.73	17.51	87.16
	GPT-3.5 (OpenAI, 2023)	6.63	4.59	13.10	33.83	34.85	90.90
	KC-VQG w/o KD	41.83	34.59	30.22	62.70	72.17	91.22
	KC-VQG w/o TA	43.49	36.10	31.10	65.10	110.21	92.61
	KC-VQG(Ours)	44.53	37.24	30.62	64.32	93.79	91.62
counting	IMVQG (Krishna et al., 2019)	0.00	0.00	3.82	5.31	5.44	84.48
	VQG-GCN (Xu et al., 2021)	5.03	1.01	7.22	10.18	9.98	87.32
	MOAG (Xie et al., 2021)	0.00	0.00	4.54	4.28	2.00	79.96
	ClipCap (Mokady et al., 2021)	6.16	3.73	17.25	24.47	9.73	85.02
	KB-VQG (Xie et al., 2022)	0.72	0.30	4.10	5.59	4.28	82.66
	GPT-3.5 (OpenAI, 2023)	13.37	9.40	19.64	40.81	65.07	92.41
	KC-VQG w/o KD	35.10	28.01	33.10	59.07	197.79	93.35
	KC-VQG w/o TA	31.25	25.73	29.20	50.41	174.96	92.34
	KC-VQG(Ours)	37.33	30.74	33.97	59.21	208.59	93.63
object recognition	IMVQG (Krishna et al., 2019)	0.98	0.00	11.26	48.82	4.35	84.25
	VQG-GCN (Xu et al., 2021)	1.60	0.00	10.29	29.24	0.75	85.81
	MOAG (Xie et al., 2021)	0.00	0.00	10.70	45.05	4.56	79.12
	ClipCap (Mokady et al., 2021)	6.36	1.75	13.59	32.92	2.09	83.75
	KB-VQG (Xie et al., 2022)	1.53	0.00	10.77	41.52	9.76	84.90
	GPT-3.5 (OpenAI, 2023)	2.20	0.87	8.82	27.90	3.38	88.80
	KC-VQG w/o KD	6.78	3.06	11.81	46.58	9.38	85.39
	KC-VQG w/o TA	6.12	1.48	12.25	47.33	2.68	85.82
	KC-VQG(Ours)	23.88	16.29	19.28	57.94	63.20	89.41

Table 1: Main automatic metrics results of baselines and our model. **Bold**: the maximum value in the column for each section.

4.5. Results and Analysis

4.5.1. Automatic Evaluation Result

We conducted experiments on both the original sampled TDIUC dataset and cross-topic scenarios for all 11 question topics conducted using the same dataset. Table 1 presents a portion of the re-

sults of the automated evaluation. In particular, we choose five cross-topic scenarios with a relatively substantial volume of test data to illustrate. Table 2 presents the automatic evaluation results of the non-cross-topic scenario. We find that:

(i) The experimental results reveal that the performance of all models is significantly lower in the

Model	BLEU-4	METEOR	BertScore
IMVQG	39.56	36.50	94.49
VQG-GCN	25.11	34.48	94.90
MOAG	19.66	31.56	88.34
ClipCap	33.61	32.20	92.68
KB-VQG	39.90	35.49	94.71
GPT-3.5	3.81	11.64	90.04
KC-VQG w/o KD	46.34	39.14	95.99
KC-VQG w/o TA	47.82	39.06	96.31
KC-VQG(Ours)	46.85	39.20	95.72

Table 2: The automatic evaluation result of non-cross-topic scenario. **Bold**: the maximum value in the column.

Model	Flu	Img_rl	Ans_rl	Top_rl
IMVQG	0.39	0.35	0.46	0.38
VQG-GCN	2.58	0.48	0.23	0.64
ClipCap	1.89	1.76	1.20	2.02
MOAG	1.25	0.89	0.35	1.01
KB-VQG	0.93	1.88	0.72	2.13
GPT-3.5	4.22	2.33	3.84	3.25
KC-VQG	3.84	3.12	2.44	3.38

Table 3: The human evaluation results in all cross-topic learning scenarios. **Bold**: the maximum value in the column.

cross-topic learning scenario compared to the non-cross-topic scenario. This demonstrates the challenging nature of our proposed CTL-VQG task. The KC-VQG model we introduced exhibits the least reduction in effectiveness in cross-topic scenarios. To be specific, in the non-cross-topic scenario, the KC-VQG model achieves a BLEU-4 score of 46.85, as indicated in Table 2. However, in the attribute cross-topic scenario, the BLEU-4 score drops to 16.63 (a relative decrease of 64.50% from 46.85 to 16.63). This contrasts with other baseline models, which experience a decrease of over 80% in attribute cross-topic learning scenarios.

(ii) Our model achieves the best results for all metrics in all cross-topic learning scenarios. Specifically, our proposed KC-VQG model achieves a BLEU-4 score of 17.24 in activity cross-topic learning scenarios, representing a significant improvement over the best-performed baseline models IMVQG (6.72) by 10.52 points. This demonstrates the capability of our proposed model to capture topic-related content, i.e., topic-related intuitive visual features and topic-related implicit knowledge.

(iii) The experimental results of the KC-VQG w/o

Model	model_cost	api_cost	total
IMVQG	38.36	-	38.36
VQG-GCN	228.81	-	228.81
ClipCap	532.89	-	532.89
MOAG	4.72	-	4.72
KB-VQG	716.28	-	716.28
KC-VQG	115.33	126.92	242.25

Table 4: The inference speed evaluation results (seconds/1000 questions). The api_cost column in the table indicates the time taken by our KC-VQG model in the Knowledge Generator module to generate knowledge. **Bold**: the minimum value in the column.

KD model and KC-VQG w/o TA model are inferior to those of our proposed KC-VQG model. This observation highlights the effectiveness of our model in accurately selecting knowledge, facilitated by the knowledge discriminator module, and the efficiency of our proposed IT-FE in extracting topic-related visual features for question generation.

(iv) The GPT-3.5 can generate suitable questions but differs from the ground truth, resulting in a lower score according to automatic evaluation metrics.

4.5.2. Human Evaluation Result

Following the research of (Xie et al., 2021), we randomly selected 200 questions generated from all cross-topic scenarios for human evaluation. We recruited 5 highly educated volunteers from a college setting to individually rate all samples. To validate the reliability of the human evaluations, we employed Fleiss Kappa Coefficients (Vieira et al., 2010) to measure both score consistency (with an average of >0.2) and score ranking consistency (with an average of >0.3). These findings indicate the credibility of our human evaluation results.

The results of the human evaluation are presented in Table 3. With the exception of the GPT-3.5 model, the human evaluation results closely align with the automatic evaluation results, showcasing the significant superiority of our model over other baseline models. Our model outperforms the GPT-3.5 model in both topic relevance and image relevance metrics. This demonstrates that the KC-VQG model we introduced excels at capturing topic-related content more effectively. Although it falls short compared to GPT-3.5 in terms of fluency and answer relevance metrics, our model can be improved by adopting a more advanced decoder (e.g., GPT-3) in the future to enhance question fluency and answer relevance.



	<p>Generated Knowledge:</p> <ol style="list-style-type: none"> 1. Glasses can be used to correct vision or protect the eyes from sunlight. 2. Trees are living organisms that provide shade, oxygen, and habitats for animals. 3. A vest is a sleeveless garment that . . . 4. A face is the front part of a person's head where . . . 5. A tie is a long, thin piece of fabric that is worn around the neck in a knot, usually as a decorative accessory. 6.No Knowledge 		<p>Generated Knowledge:</p> <ol style="list-style-type: none"> 1. There is one girl in the image. 2. The girl is wearing glasses. 3. There is one tree in the background. 4. The girl is wearing one pink dress. 5. The girl is playing with one frisbee. 6. No Knowledge
<p>Topic: attribute</p>		<p>Topic: counting</p>	
<p>Generated Questions:</p> <p>KC-VQG (ours) : What material is the tie shown in the picture? <input checked="" type="checkbox"/></p> <p>IMVQG : What is to man doing doing the?</p> <p>VQG-GCN : What is the man doing?</p> <p>CLIPCap : What is he wearing what is he doing at the office?</p> <p>MOAG : what is behind behind the?</p> <p>KB-VQG : How is the man doing doing the?</p> <p>Ground Truth : What is the vest made of?</p>		<p>Generated Questions:</p> <p>KC-VQG (ours) : How many cars are in the photo? <input checked="" type="checkbox"/></p> <p>IMVQG : What is is the?</p> <p>VQG-GCN : What are the people doing?</p> <p>CLIPCap : How is her wearing around her neck most important?</p> <p>MOAG : What is the the the?</p> <p>KB-VQG : What is is the any?</p> <p>Ground Truth : How many cars are there?</p>	
(a)		(b)	

Figure 4: Case study of generated questions by our model and baseline models. The knowledge highlighted in red in the figure represents the most suitable knowledge chosen by the knowledge discriminator.

4.5.3. Inference Speed Evaluation

In order to compare the inference speed of our proposed KC-VQG model with various baseline models in generating questions, we conduct experiments on 1000 test samples for each model in the same experimental environment with a single GTX2080 Ti GPU. The results of the model efficiency evaluation are presented in Table 4.

The experiment shows that our model exhibits faster generation times than the KB-VQG and Clip-Cap models, slower generation times than the IMVQG and MOAG models, and comparable generation times to the VQG-GCN model. While our model lag behind the simple and straightforward IMVQG and MOAG models in inference speed, it significantly outperforms them in generating questions of superior quality in cross-topic scenarios.

4.6. Case Study

Fig. 4 shows several examples of the generated question by our KC-VQG model and baseline models. As depicted in Figure 4 (a) and (b), it is evident that the outcomes produced by all baseline models exhibit a weak correlation with the topic. This deficiency arises due to the baseline model's inability to establish a meaningful connection between the unseen topic and the image, further compounded by the absence of topic-related knowledge as supplementary input. In contrast, our model, equipped with the IT-KE and the IT-FE module, addresses this limitation effectively. As shown in Fig. 4 (a) and (b), our proposed KC-VQG model demonstrates the capability to select relevant knowledge based on the input topic and produce coherent, topic-related questions within cross-topic scenarios.

5. Conclusion

In this paper, we introduce a new task, i.e., the cross-topic learning visual question generation (CTL-VQG) task. This task aims to generate questions that are learned within a limited number of labeled topics and effectively extends to those unlabeled topics. To solve this CTL-VQG task, we propose to capture content related to a new topic without annotation while maintaining structural information. We propose an image-topic feature extractor to extract topic-related intuitive visual features for a given image. We propose an image-topic knowledge extractor to extract and select the most appropriate topic-related implicit knowledge for generating questions from large language models (LLMs). We propose a topic and type-guided question decoder to generate questions based on topic-related information and question type. Experimental results demonstrate that our proposed model outperforms other baselines on all cross-topic learning scenarios, which achieves state-of-the-art performance on this CTL-VQG task.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (62076100), Fundamental Research Funds for the Central Universities, SCUT (x2rjD2230080), the Science and Technology Planning Project of Guangdong Province (2020B0101100002), Guangdong Provincial Fund for Basic and Applied Basic Research - Regional Joint Fund Project (Key Project) (23201910250000318,308155351064), CAAI-Huawei MindSpore Open Fund, CCF-Zhipu AI Large Model Fund.

7. Bibliographical References

- Chao Bi, Shuhui Wang, Zhe Xue, Shengbo Chen, and Qingming Huang. 2022. Inferential visual question generation. In *Proc. of ACM MM*, page 4164–4174.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proc. of NeurIPS*.
- Wei-Lun Chao, Hexiang Hu, and Fei Sha. 2018. Cross-dataset adaptation for visual question answering. In *Proc. of CVPR*.
- Feng Chen, Jiayuan Xie, Yi Cai, Tao Wang, and Qing Li. 2021. Difficulty-controllable visual question generation. In *Proc. of APWeb-WAIM*, pages 332–347.
- Jiali Chen, Zhenjun Guo, Jiayuan Xie, Yi Cai, and Qing Li. 2023. Deconfounded visual question generation with causal inference. In *Proc. of ACM MM*, page 5132–5142.
- Shiming Chen, Ziming Hong, Guo-Sen Xie, Wenhao Yang, Qinmu Peng, Kai Wang, Jian Zhao, and Xinge You. 2022. MSDN: mutually semantic distillation network for zero-shot learning. In *Proc. of CVPR*, pages 7602–7611.
- Michael J. Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proc. of ACL Workshop*, pages 376–380.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pages 4171–4186.
- James Thomas Dillon. 2004. *Questioning and teaching: A manual of practice*.
- Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. 2022. Decoupling zero-shot semantic segmentation. In *Proc. of CVPR*, pages 11573–11582.
- Huizhong Duan, Yunbo Cao, Chin-Yew Lin, and Yong Yu. 2008. Searching questions by identifying question topic and question focus. In *Proc. of ACL*, pages 156–164.
- Sepideh Esmailpour, Bing Liu, Eric Robertson, and Lei Shu. 2022. Zero-shot out-of-distribution detection based on the pre-trained model CLIP. In *Proc. of AAAI*, pages 6568–6576.
- Z. Fan, Zhongyu Wei, P. Li, Y. Lan, and X. Huang. 2018. A question type driven framework to diversify visual question generation. In *Proc. of IJCAI*.
- Zhiyu Fang, Xiaobin Zhu, Chun Yang, Zheng Han, Jingyan Qin, and Xu-Cheng Yin. 2022. Learning aligned cross-modal representation for generalized zero-shot classification. In *Proc. of AAAI*, pages 6605–6613.
- Ariel Gera, Alon Halfon, Eyal Shnarch, Yotam Perlitz, Liat Ein-Dor, and Noam Slonim. 2022. Zero-shot text classification with self-training. In *Proc. of EMNLP*, pages 1107–1119.
- Bin He, Meng Xia, Xinguo Yu, Pengpeng Jian, Hao Meng, and Zhanwen Chen. 2017. An educational robot system of visual question answering for preschoolers. In *proc. of ICRAE*, pages 441–445.
- Peiliang Huang, Junwei Han, De Cheng, and Dingwen Zhang. 2022. Robust region feature synthesizer for zero-shot object detection. In *Proc. of CVPR*, pages 7612–7621.
- Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. 2022. Zero-shot text-guided object generation with dream fields. In *Proc. of CVPR*, pages 857–866.
- Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *Proc. of ICCV*.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.
- Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. 2019. Information maximizing visual question generation. In *Proc. of CVPR*, pages 2008–2018.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proc. of ACL*, pages 7871–7880.

- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. of ACL Workshop*, page 74–81.
- Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. Clipcap: CLIP prefix for image captioning. *CoRR*, abs/2111.09734.
- Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proc. of ACL*.
- OpenAI. 2023. [Chatgpt \(september 25 version\) \[large language model\]](#).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proc. of ICML*, volume 139, pages 8748–8763.
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training.
- 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):9.
- Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. 2022. Nlx-gpt: A model for natural language explanations in vision and vision-language tasks. In *Proc. of CVPR*, pages 8322–8332.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. [Conceptnet 5.5: An open multilingual graph of general knowledge](#). pages 4444–4451.
- Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proc. of CVPR*, pages 17897–17907.
- Jialin Tian, Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen. 2022. TVT: three-way vision transformer through multi-modal hypersphere learning for zero-shot sketch-based image retrieval. In *Proc. of AAAI*, pages 2370–2378.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NeurIPS*, pages 5998–6008.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proc. of CVPR*, pages 4566–4575.
- Nihir Vedd, Zixu Wang, Marek Rei, Yishu Miao, and Lucia Specia. 2022. Guiding visual question generation. In *Proc. of NAACL*, pages 1640–1654.
- Susana M Vieira, Uzay Kaymak, and João MC Sousa. 2010. Cohen’s kappa coefficient as a performance measure for feature selection. In *International conference on fuzzy systems*, pages 1–8. IEEE.
- Jiayuan Xie, Yi Cai, Qingbao Huang, and Tao Wang. 2021. Multiple objects-aware visual question generation. In *Proc. of ACM MM*, pages 4546–4554.
- Jiayuan Xie, Wenhao Fang, Yi Cai, Qingbao Huang, and Qing Li. 2022. Knowledge-based visual question generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7547–7558.
- Xing Xu, Tan Wang, Yang Yang, Alan Hanjalic, and Heng Tao Shen. 2021. Radial graph convolutional network for visual question generation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(4):1654–1667.
- Mingda Zhang, Tristan Maidment, Ahmad Diab, Adriana Kovashka, and Rebecca Hwa. 2021. Domain-robust vqa with diverse datasets and methods but no target labels. In *Proc. of CVPR*, pages 7046–7056.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTscore: Evaluating text generation with BERT. In *Proc. of ICLR*.
- Ye Zheng, Jiahong Wu, Yongqiang Qin, Faen Zhang, and Li Cui. 2021. Zero-shot instance segmentation. In *Proc. of CVPR*, pages 2593–2602.