様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 １４ 日現在

機関番号：１２６０１
研究種目：若手研究
研究期間：2022～2023
課題番号：２２Ｋ１７９４７

研究課題名（和文）Vision and language cross-modal for training conditional GANs with long-tail data.

研究課題名（英文）Vision and language cross-modal for training conditional GANs with long-tail data.

研究代表者

ヴォ　ミンデュク（VO, MinhDuc）

東京大学・大学院情報理工学系研究科・特任助教

研究者番号：４０９３９９０６

交付決定額（研究期間全体）：（直接経費）　　2,000,000円

研究成果の概要（和文）：本研究は、視覚と言語の空間間におけるクロスモダリティに関する知識を得ることを目的としています。私たちは、物体の視覚的外観と対応する言語記述を含む知識ベースを構築しました。収集された知識ベースが、見たことのない物体の記述能力を向上させ、未来を予測する能力を強化することを実証しました。

また、限られたデータセットやオープンセットデータセットの下での生成的敵対的ネットワーク（GAN）のトレーニングおよびGANインバージョンの新しいトレーニングパラダイムを探求しました。

研究成果の学術的意義や社会的意義
We shows the efficacy of external knowledge base, helping AI in understanding up-to-date object knowledge and being able to predict the future given a sequence of sparsely temporally-ordered images. We showed the ability of generative AI when it is trained using limited number of training data.

研究成果の概要（英文）：This study gains the knowledge about cross-modality between vision and language spaces. We built the knowledge base containing object's visual appearance and corresponding language description. We illustrated the efficacy of the collected knowledge base in enhancing the ability of describing unseen objects and predicting the future.
We also explored new training paradigms of training generative adversarial networks under limited and open-set dataset as well as GAN inversion. This illustrated the ability of training a generative model when we cannot always harvest enough data to train a generative AI.

研究分野：Computer vision

キーワード：Vision and language　Novel object captioning　GANs　External knowledge

様　式　C－１９、F－１９－１（共通）

１．研究開始当初の背景

Large language models (LLMs)-based image captioning has the capability of describing objects not explicitly observed in training data; yet novel objects occur frequently, necessitating the requirement of sustaining up-to-date object knowledge for open-world comprehension.

２．研究の目的

We aim to build a highly effective image captioning method that can easily be adapted to ever-changing object knowledge.

３．研究の方法

Instead of relying on large amounts of data and/or scaling up network parameters, we introduce a highly effective retrieval-augmented image captioning method that prompts LLMs with object names retrieved from external visual-name memory (EVCAP). We build ever-changing object knowledge memory using objects' visuals and names, enabling us to (i) update the memory at a minimal cost and (ii) effortlessly augment LLMs with retrieved object names by utilizing a lightweight and fast-to-train model
Our method involves two challenges: (1) building an external memory containing up-to-date objects, and (2) building an effective LLMs-based model using retrieved object names.
For (1), we first collect image-name pairs from the external data source where the image can be either real or synthesized. After that, we encode these images into image embeddings, which serve as keys in memory, and use their names as values.
For (2), we build a model consisting of three key components (Fig. 1): object name retrieval to retrieve object' names from the external memory by matching between object's features and those in the memory; attentive fusion to eliminate redundant objects' names; and caption generation to prompt a pre-trained LLM using retrieved object's names.

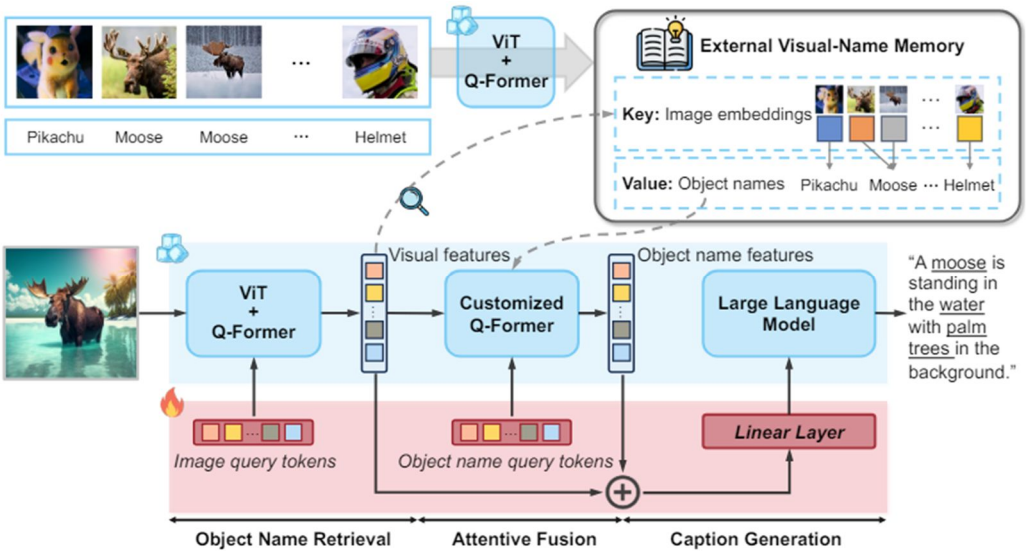

Figure 2. Schematic of our proposed EVCAP. It consists of an external visual-name memory with image embeddings and object names (upper), a frozen ViT and Q-Former equipped with trainable image query tokens, an attentive fusion module developed by a customized frozen Q-Former and trainable object name query tokens, and a frozen LLM with a trainable linear layer (lower). The ViT and QFormer extract learned visual features from the input image, which are then used to retrieve object names from the external memory. These retrieved object names and learned visual features undergo cross-attention in the customized Q-Former, creating refined object name features. Finally, the object name features combined with visual features are fed into the LLM post a linear layer for generating captions.

4. 研究成果

Our model, which was trained only on the COCO dataset, can adapt to out-of-domain without requiring additional fine-tuning or re-training. Our experiments conducted on benchmarks including COCO, Nocap, Flickr30k, and synthetic commonsense-violating data. In comparison with SOTA methods, Table 1 details our EVCAP's performance on in-/out-domain benchmarks and Table 2 shows the results on commonsense-violating data. These results show that EV-CAP, with only 3.97M trainable parameters, exhibits superior performance compared to other methods based on frozen pre-trained LLMs. Its performance is also competitive compared to specialist SOTAs that require extensive training.

Table 1. Quantitative comparison against SOTA methods on three common image captioning benchmarks. * denotes using a memory bank. We report the size of training data and parameters; BLEU@4 (B@4), METEOR (M), CIDEr (C), and SPICE (S) scores on COCO test set; C and S scores on in-domain, near-domain, out-domain and overall data of NoCaps validation set; C and S scores on Flickr30k test set. Higher score is better. Bold indicates the best results among compared methods, normal indicates the second best results.

Method	Training		COCO Test				NoCaps val								Flickr30k Test	
	Data	Para.					In-domain		Near-domain		Out-domain		Overall			
			B@4	M	C	S	C	S	C	S	C	S	C	S	C	S
Heavyweight-training models																
VinVL [45]	8.9M	110M	38.2	30.3	129.3	23.6	96.8	13.5	90.7	13.1	87.4	11.6	90.9	12.8	–	–
AoANet+MA* [16]	COCO	–	38.0	28.7	121.0	21.8	–	–	–	–	–	–	–	–	–	–
NOC-REK* [40]	COCO	110M	–	–	–	–	104.7	14.8	100.2	14.1	100.7	13.0	100.9	14.0	–	–
RCA-NOC* [13]	COCO	110M	37.4	29.6	128.4	23.1	92.2	12.9	87.8	12.6	87.5	11.5	88.3	12.4	–	–
ViECap $_{GPT2}$ [15]	COCO	124M	27.2	24.8	92.9	18.2	61.1	10.4	64.3	9.9	65.0	8.6	66.2	9.5	47.9	13.6
InstructBLIP $_{Vicuna-13B}$ [11]	129M	188M	–	–	–	–	–	–	–	–	–	–	121.9	–	82.8	–
OSCAR [26]	4.1M	338M	37.4	30.7	127.8	23.5	83.4	12.0	81.6	12.0	77.6	10.6	81.1	11.7	–	–
BLIP [24]	129M	446M	40.4	–	136.7	–	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	–	–
BLIP-2 $_{FlanT5-XL}$ [25]	129M	1.2B	**42.4**	–	144.5	–	**123.7**	**16.3**	**120.2**	**15.9**	**124.8**	**15.1**	121.6	**15.8**	–	–
REVEAL* $_{T5}$ [20]	1.3B	2.1B	–	–	**145.4**	–	–	–	–	–	–	–	**123.0**	–	–	–
Lightweight-training models																
MiniGPT4 $_{Vicuna-13B}$ [46]	5M	**3.94M**	38.0	29.6	129.6	23.4	99.0	14.8	106.9	15.3	110.8	14.9	108.8	15.1	78.4	16.9
SmallCap* $_{GPT2}$ [35]	COCO	7M	37.0	27.9	119.7	21.3	–	–	–	–	–	–	–	–	60.6	–
ClipCap $_{GPT2}$ [29]	COCO	43M	33.5	27.5	113.1	21.1	84.9	12.1	66.8	10.9	49.1	9.6	65.8	10.9	–	–
EVCAP* $_{Vicuna-13B}$	COCO	3.97M	41.5	**31.2**	140.1	**24.7**	111.7	15.3	119.5	15.6	116.5	14.7	119.3	15.3	**84.4**	**18.0**
Specialist SOTAs																
Qwen-VL $_{Qwen-7B}$ [5]	1.4B	9.6B	–	–	–	–	–	–	–	–	–	–	121.4	–	85.8	–
CogVLM $_{Vicuna-7B}$ [41]	1.5B	6.5B	–	–	148.7	–	–	–	–	–	132.6	–	128.3	–	94.9	–
PaLI $_{mT5-XXL}$ [9]	1.6B	17B	–	–	149.1	–	–	–	–	–	–	–	127.0	–	–	–
PaLI-X $_{UL2-32B}$ [8]	2.2B	55B	–	–	149.2	–	–	–	–	–	–	–	126.3	–	–	–

Table 2. Quantitative results on commonsense-violating data - WHOOPS dataset. EVCAP (w/ WHOOPS) denotes EVCAP using the memory expanded by WHOOPS objects. The results reveal the open-world comprehension ability and expandability of EVCAP.

Method	B@4	M	C	S
Only pre-trained models				
BLIP [24] (from [6])	13	–	65	–
BLIP-2 $_{FlanT5-XXL}$ [25] (from [6])	31	–	120	–
BLIP-2 $_{FlanT5-XXL}$ [25] (reproduced)	28	26.7	93.1	17.9
Finetuned models on COCO				
MiniGPT4 [46]	24.2	26.7	84.8	18.2
BLIP [24]	22.9	25.0	79.3	17.1
BLIP-2 $_{FlanT5-XL}$ [25]	25.8	27.0	89.1	18.3
End-to-end trained models on COCO				
EVCAP	24.1	26.1	85.3	17.7
EVCAP (w/ WHOOPS)	24.4	26.1	86.3	17.8

Figure 1 presents a comparison of captions generated by our EVCAP and three SOTA models across three benchmarks.

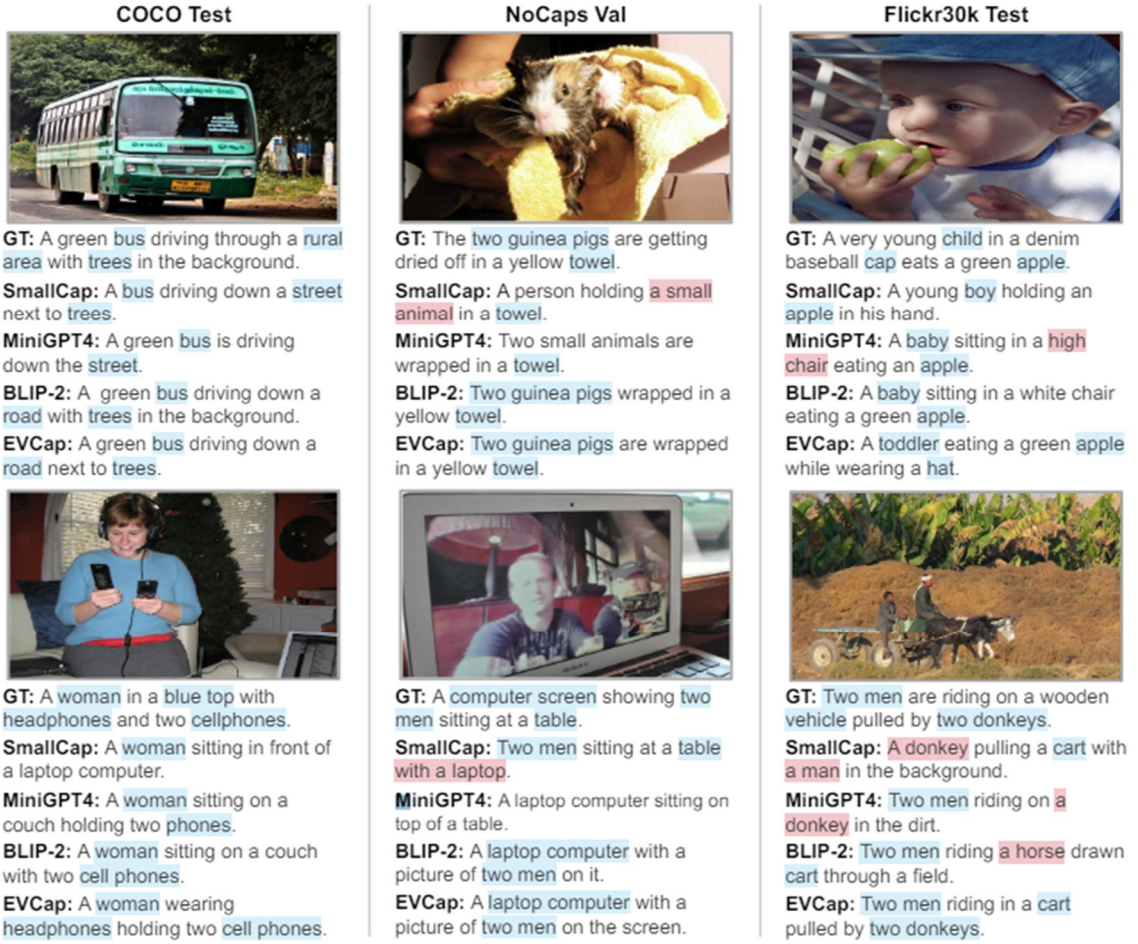

Figure 3. Examples of captions generated by our EVCAP and three SOTA methods on COCO test set, NoCaps validation set, and Flickr30k test set. GT refers to the Ground Truth captions. Incorrect objects in captions are highlighted in red , while correct ones are in blue . Our EVCAP correctly generates captions across different datasets, showing performance comparable to BLIP-2.

To explore the influence of different LLMs decoders on our EVCAP, we experiment by substituting Vicuna-13B with GPT2 and Vicuna-7B, as detailed in Table 3. Their comparison also underscores the effectiveness of our method's object name retrieval and attentive fusion strategy.

Table 3. Analysis with different LLM decoders including GPT2, Vicuna-7B, and Vicuna-13B. The results reveal EVCAP is effective when applying it in different LLM decoders.

Method	LLM	COCO test		NoCaps val		Flickr30k test	
		C	S	C	S	C	S
SmallCap [35]	GPT2	119.7	21.3	–	–	60.6	–
ViECap [15]	GPT2	92.9	18.2	66.2	9.5	47.9	13.6
EVCAP	GPT2	131.0	23.2	97.6	13.3	70.6	16.1
MiniGPT4 [46]	Vicuna-7B	119.4	23.5	108.7	15.7	73.9	17.2
InstructBLIP [11]	Vicuna-7B	–	–	123.1	–	82.4	–
EVCAP	Vicuna-7B	139.0	24.7	116.8	15.3	82.7	18.0
MiniGPT4 [46]	Vicuna-13B	129.6	23.4	108.8	15.1	78.4	16.9
InstructBLIP [11]	Vicuna-13B	–	–	121.9	–	82.8	–
EVCAP	Vicuna-13B	140.1	24.7	119.3	15.3	84.4	18.0

Though our method achieves notable performance, our method remains limitations. First, EVCAP cannot retrieve all objects that appear in the given image due to the memory coverage limits, leading to incomplete image descriptions. Second, our focus on object representation restricts consideration of other crucial captioning elements, affecting overall performance. Similar to all models trained with COCO dataset, EVCAP has limitations in generating varied styles.

５．主な発表論文等

〔雑誌論文〕　計10件（うち査読付論文　10件／うち国際共著　10件／うちオープンアクセス　10件）

１．著者名 Duc Minh Vo, Hong Chen, Akihiro Sugimoto, Hideki Nakayama	４．巻 -
２．論文標題 NOC-REK: Novel Object Captioning with Retrieved Vocabulary from External Knowledge	５．発行年 2022年
３．雑誌名 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)	６．最初と最後の頁 17979 - 17987
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/CVPR52688.2022.01747	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Rui Yang, Duc Minh Vo, Hideki Nakayama	４．巻 10
２．論文標題 Stochastically Flipping Labels of Discriminator's Outputs for Training Generative Adversarial Networks	５．発行年 2022年
３．雑誌名 IEEE Access	６．最初と最後の頁 103644 - 103654
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2022.3210130	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Hong Chen, Duc Vo, Hiroya Takamura, Yusuke Miyao, Hideki Nakayama	４．巻 -
２．論文標題 StoryER: Automatic Story Evaluation via Ranking, Rating and Reasoning	５．発行年 2022年
３．雑誌名 Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing	６．最初と最後の頁 1739 - 1753
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Katsumata Kai、Vo Duc Minh、Harada Tatsuya、Nakayama Hideki	４．巻 1
２．論文標題 Soft Curriculum for Learning Conditional GANs with Noisy-Labeled and Uncurated Unlabeled Data	５．発行年 2024年
３．雑誌名 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)	６．最初と最後の頁 5311-5320
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/WACV57701.2024.00524	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Katsumata Kai、Vo Duc Minh、Liu Bei、Nakayama Hideki	４．巻 1
２．論文標題 Revisiting Latent Space of GAN Inversion for Robust Real Image Editing	５．発行年 2024年
３．雑誌名 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)	６．最初と最後の頁 5301-5310
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/WACV57701.2024.00523	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Katsumata Kai、Vo Duc Minh、Nakayama Hideki	４．巻 1
２．論文標題 Label Augmentation as Inter-class Data Augmentation for Conditional Image Synthesis with Imbalanced Data	５．発行年 2024年
３．雑誌名 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)	６．最初と最後の頁 4932-4941
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/WACV57701.2024.00487	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Li Jiaxuan、Vo Duc Minh、Nakayama Hideki	４．巻 1
２．論文標題 Partition-and-Debias: Agnostic Biases Mitigation via A Mixture of Biases-Specific Experts	５．発行年 2023年
３．雑誌名 2023 IEEE/CVF International Conference on Computer Vision (ICCV)	６．最初と最後の頁 4901-4911
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ICCV51070.2023.00454	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Vo Duc Minh、Luong Quoc-An、Sugimoto Akihiro、Nakayama Hideki	４．巻 1
２．論文標題 A-CAP: Anticipation Captioning with Commonsense Knowledge	５．発行年 2023年
３．雑誌名 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)	６．最初と最後の頁 10824-10833
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/CVPR52729.2023.01042	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 Yang Rui、Vo Duc Minh、Nakayama Hideki	4．巻 1
2．論文標題 Indirect Adversarial Losses via an Intermediate Distribution for Training GANs	5．発行年 2023年
3．雑誌名 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV	6．最初と最後の頁 4641-4650
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/WACV56688.2023.00463	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 Li Jiaxuan、Vo Duc Minh、Sugimoto Akihiro、Nakayama Hideki	4．巻 1
2．論文標題 EVCap: Retrieval-Augmented Image Captioning with External Visual--Name Memory for Open-World Comprehension	5．発行年 2024年
3．雑誌名 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計9件（うち招待講演　0件／うち国際学会　9件）

1．発表者名 Duc Minh Vo, Hong Chen, Akihiro Sugimoto, Hideki Nakayama
2．発表標題 NOC-REK: Novel Object Captioning with Retrieved Vocabulary from External Knowledge
3．学会等名 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)（国際学会）
4．発表年 2022年

1．発表者名 Hong Chen, Duc Vo, Hiroya Takamura, Yusuke Miyao, Hideki Nakayama
2．発表標題 StoryER: Automatic Story Evaluation via Ranking, Rating and Reasoning
3．学会等名 2022 Conference on Empirical Methods in Natural Language Processing（国際学会）
4．発表年 2022年

１．発表者名
Katsumata Kai、Vo Duc Minh、Harada Tatsuya、Nakayama Hideki

２．発表標題
Soft Curriculum for Learning Conditional GANs with Noisy-Labeled and Uncurated Unlabeled Data

３．学会等名
2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV（国際学会）

４．発表年
2024年

１．発表者名
Katsumata Kai、Vo Duc Minh、Liu Bei、Nakayama Hideki

２．発表標題
Revisiting Latent Space of GAN Inversion for Robust Real Image Editing

３．学会等名
2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV（国際学会）

４．発表年
2024年

１．発表者名
Katsumata Kai、Vo Duc Minh、Nakayama Hideki

２．発表標題
Label Augmentation as Inter-class Data Augmentation for Conditional Image Synthesis with Imbalanced Data

３．学会等名
2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV（国際学会）

４．発表年
2024年

１．発表者名
Li Jiaxuan、Vo Duc Minh、Nakayama Hideki

２．発表標題
Partition-and-Debias: Agnostic Biases Mitigation via A Mixture of Biases-Specific Experts

３．学会等名
2023 IEEE/CVF International Conference on Computer Vision (ICCV（国際学会）

４．発表年
2023年

１．発表者名
Vo Duc Minh、Luong Quoc-An、Sugimoto Akihiro、Nakayama Hideki

２．発表標題
A-CAP: Anticipation Captioning with Commonsense Knowledge

３．学会等名
2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR（国際学会）

４．発表年
2023年

１．発表者名
Yang Rui、Vo Duc Minh、Nakayama Hideki

２．発表標題
Indirect Adversarial Losses via an Intermediate Distribution for Training GANs

３．学会等名
2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV（国際学会）

４．発表年
2023年

１．発表者名
Li Jiaxuan、Vo Duc Minh、Sugimoto Akihiro、Nakayama Hideki

２．発表標題
EVCap: Retrieval-Augmented Image Captioning with External Visual--Name Memory for Open-World Comprehension

３．学会等名
2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)（国際学会）

４．発表年
2024年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関