Fourth-person Captioning: Describing Daily Events by Uni-supervised and Tri-regularized Training

Kazuto Nakashima
Graduate School of ISEE
Kyushu University
Fukuoka, Japan
k_nakashima@irvs.ait.kyushu-u.ac.jp

Yumi Iwashita
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA, USA
Yumi.Iwashita@jpl.nasa.gov

Akihiro Kawamura, Ryo Kurazume
Department of ISEE
Kyushu University
Fukuoka, Japan
{kawamura, kurazume}@ait.kyushu-u.ac.jp

***Abstract*—We aim to develop a supporting system which enhances the ability of human's short-term visual memory in an intelligent space where the human and a service robot coexist. Particularly, this paper focuses on how we can interpret and record diverse and complex life events on behalf of humans, from a multi-perspective viewpoint. We propose a novel method named "fourth-person captioning", which generates natural language descriptions by summarizing visual contexts complementarily from three types of cameras corresponding the first-, second-, and third-person viewpoint. We first extend the latest image captioning technique and design a new model to generate a sequence of words given the multiple images. Then we provide an effective training strategy that needs only annotations supervising images from a single viewpoint in a general caption dataset and unsupervised triplet instances in the intelligent space. As the three types of cameras, we select a wearable camera on the human, a robot-mounted camera, and an embedded camera, which can be defined as the first-, second-, and third-person viewpoint, respectively. We hope our work will accelerate a cross-modal interaction bridging the human's egocentric cognition and multi-perspective intelligence.**

***Index Terms*—Intelligent space, image captioning**

Fig. 1: Examples of the first-person images (left column), the second-person images (middle column), and third-person images (right column) in the intelligent space.

I. Introduction

Intelligent space, the room or the area that is equipped with various sensors or cameras, has been widely studied in the robotics community because of its feasibility of human-robot coexistence [1]–[3]. Although it is difficult for a stand-alone robot to observe the dynamic environment and operate diverse service tasks for the humans with only on-board sensors, intelligent space enables it to expand its observation area. Moreover, the studies on intelligent space aim not only to compensate their sensing ability but also to process the daily streamed data and explore the information structure on behalf of the robot [2], [3].

Here we address the problem in which how the intelligent space can be used to augment the human's short-term visual memory. Some studies [2], [3] focused on the human-centered system as an external memory in the intelligent space. Niitsuma *et al.* [2] proposed a user interface which enables humans to virtually arrange the computerized information to the real 3D space and to retrieve them by pointing. In our previous work [3], we developed a glasses-type interface where the users can know about the household items managed by the intelligent space, by passing object keywords or just looking the shelves. Another approach to achieve the augmented visual memories is to pool the autobiographical videos in a daily life, extract the latent semantics, and configure a video retrieval system with a high-affinity interface for humans [4]. Particularly, the first-person videos captured from the user's egocentric viewpoint have been attracting increasing interests for that purpose. Miyanishi *et al.* [5] proposed a gesture-based video retrieval system. Nakayama *et al.* [6] proposed a goggle-type system which retrieves the past images with object label queries. Meanwhile, Fan *et al.* [7] adopted the encoder-decoder neural network [8] to generate natural language description for the first-person image and developed a retrieval system. The natural language sentence given to the image explains not only the related objects but also the relationship between objects. Compared to the gestures and the object labels, it is literally a *natural* way for humans to generate flexible queries, *e.g.*, they can easily contain temporal terms. Due to the nature that the first-person videos contain undergoing hand manipulations and saliency objects, the first-person videos have been used for not only the automated lifelogging [5]–[7] but also wearer's action recognition [9], [10] and so on. The first-person viewpoint can now easily be acquired by wearable devices, however, the view field is narrow and visual information about the wearer's

surrounding fragmentarily appears.

Thus we focus on the multi-perspective vision we can leverage in the intelligent space. In the human-robot coexistence situation, not just the user's egocentric viewpoint, a robot-mounted camera to monitor the human closely can also be used. We define it as the second-person viewpoint. Moreover, the intelligent space generally has embedded cameras to observe the comprehensive state. We define this type of camera as the third-person viewpoint. In our previous work [11], we proposed "the fourth-person viewpoint" which complementarily integrates three type of viewpoints. We assumed that each of these has unique visual concepts and can be merged to compensate incomplete information. It aimed to interpret daily events precisely like a book reader who picks up multi-perspective sentences and appreciates the story.

In this paper, we present a novel method to describe daily events with a natural language by using the fourth-person viewpoint in the intelligent space. The triplet images via the fourth-person viewpoint have unique resolutions and unique visible objects, which may raise the likelihood of the precise caption generation. To our knowledge, we are the first to generate image captions from the multi-perspective viewpoints; the first-, second-, and third-person viewpoints.

II. Related Work

In this section, we describe relevant research background.

Most researches on the fusion of the multi-perspective images focused on labeling of co-occurrence events related to multiple viewpoints. Yonetani *et al.* [10] improved the recognition of interactive actions/reaction from a pair of the first- and second-person viewpoints. Fan *et al.* [12] proposed a method to identify the first-person video from some camera wearers in the third-person videos. Other works leveraged the multiple videos for cross-view action recognition [13] and event detection [14]. These studies focused on the classification/identification of human actions and have not reached a comprehensive description including handling objects and the context.

On the other hand, in recent years, the task of generating image captions expressing visual information with natural language sentences has developed rapidly, motivated by advances in deep neural networks [7], [8], [15]. Image captioning is the primary form of understanding real-world scenes and is indispensable for a high-affinity interface in the future intelligent systems. However, conventional image caption research did not explicitly consider the modality by *person*, and handled images of all viewpoints in a single input model. Some studies [7], [15] focused on the first-person images/videos for image captioning task. The work by Fan et al. [7] is most closely related to ours. They generated a caption from the first-person image sequences, in hopes of automatic lifelogging. They gathered the first-person images from a wearable camera, construct an annotated dataset, and achieved to generate the "first-person" sentences. However, it would take too much cost to newly annotate images. Even the work [7] above used a common dataset which is composed of random images. It can be still more a crucial problem for our muti-perspective situation. Therefore, we tackle the problem by taking a strategy to solve our objective with only single-perspective supervision.

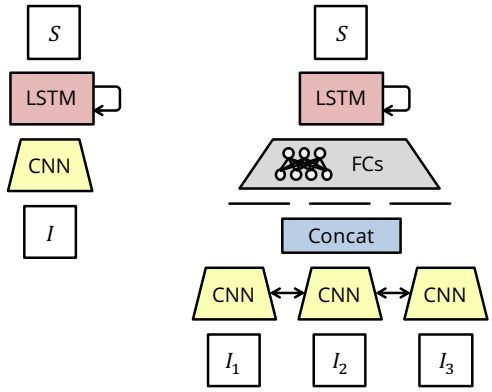


Fig. 2: A basic architecture for generating a sequence of words S given a single image I (left) [8] and naïve extension for multiple inputs I_1, I_2, I_3 (right). The "CNN", "LSTM", "FCs", and "Concat" denote a convolutional neural network, a long shot-term memory module, fully-connected layers, and concatenating operation, respectively.

III. Captioning Model

In this section, we describe the models that we use for the caption generation from three types of images. Our model is extended version of the encoder-decoder model [8] as shown in Fig. 2. We add a fully-connected fusion layer which projects the triplet image features produced by the encoder to a single representation vector.

A. Image Encoder

Our model takes three types of raw images from a wearable camera, a robot, and an embedded camera. We use a convolutional neural network (CNN) pre-trained on the large-scale image dataset such as ImageNet [16] in order to encode them as a set of *viewpoint-wise* feature vectors f_i, each of which is a D_1-dimensional global image representation.

B. Fusion Layer

The triplet image features $f_i \in \mathbb{R}^{D_1}$ are simply concatenated and fed into a fully-connected layer to project to a multimodal space $\mathbb{R}^{D_2}$ which is followed by the caption decoder.

$$f' = W \cdot f \tag{1}$$

$$f = [f_1, f_2, f_3]^{\mathrm{T}} \tag{2}$$

where $W \in \mathbb{R}^{D_2 \times 3D_1}$ is learned parameters. The f' is followed by a non-linearity and Batch Normalization [17]. We employ a hyperbolic tangent for the non-linearity.

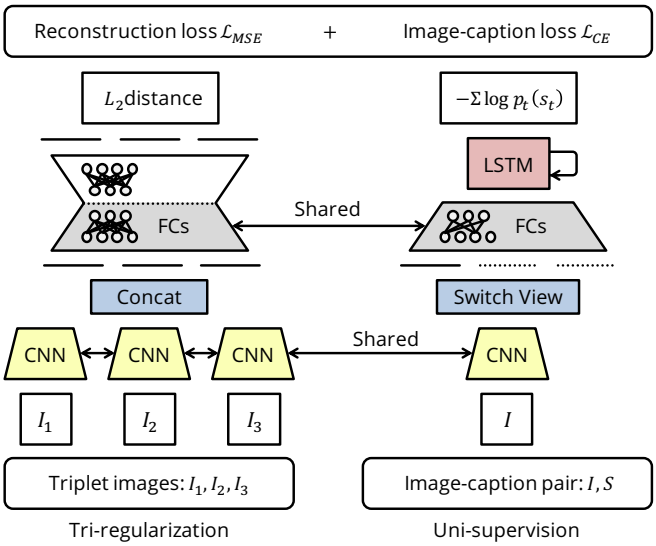


Fig. 3: Schematic illustration of training of our multi-perspective captioning model. In the uni-supervision stream, the visual information (solid lines) is fed into the LSTM cell once at the first time step.

C. Recurrent Caption Decoder

We use a recurrent network with a long short-term memory cell (LSTM) to generate words of a caption $S = \{s_0, \ldots, s_T\}$. At each time step t, the LSTM updates its hidden state h_t, given previous hidden state h_{t-1} and an input vector x_t.

$$h_t = \mathrm{LSTM}(h_{t-1}, x_t) \tag{3}$$

The multi-perspective feature f' generated by the fusion layer is fed into the LSTM as x_0 at the first step $t = 0$. At the following time steps $t \geq 1$, a word s_{t-1} is embedded to a D_2-dimensional space as x_t.

$$x_t = W_e \cdot s_{t-1} \tag{4}$$

where $W_e \in \mathbb{R}^{D_2 \times D_s}$ is learned parameters. Note that a $s_t \in \mathbb{R}^{D_s}$ is a word vector in one-hot representation and its sequence starts with a special start token s_0. The state h_t is then multiplied by additional parameters W_s and fed to $\mathrm{softmax}(\cdot)$ which produces a probability distribution p_t over the pre-defined vocabulary words. The next word s_t is defined by $\mathrm{argmax}(p_t)$.

IV. Uni-supervised and Tri-regularized Training

In this section, we address the problem on training the multi-perspective model in our situation and describe our approach.

Traditionally, given an image I and a target word sequence $S = \{s_0, \ldots, s_T\}$ as groud truth, the model θ is trained by minimizing the sum of the crossentropy loss of the given words, which equals to maximize the likelihood of the target sequence.

$$\mathcal{L}_{CE} = -\sum_{t=1}^{T} \log p_t(s_t) \tag{5}$$

To compute the loss, we need to initialize the LSTM with the instance of the fused image feature at the first step, as described in the previous section. Therefore, for such an end-to-end manner, we have to newly prepare a large-scale dataset consists of pairs of the triplet image inputs and the correspoding caption outputs. However, it can be considered that the daily living environments are diverse regarding household items or room layouts and it would take too much cost to construct a new dataset which enables to learn visual-semantic relationships and generalize every case. In this paper, instead, we propose a novel training strategy to train the multi-perspective captioning model with a single perspective supervision generally used to train captioning models.

To supervise the multi-perspective model with a general pair of a single image and caption candidates, we have several options. The first approach is to duplicate the single image to three inputs for the fusion layer and train the entire model in the usual manner. This approach can sufficiently train the language model but it may overfit to the "duplicated input" situation which is totally different from the multi-perspective vision. The second approach is to pool the image features, for example by averaging or taking max values over the three viewpoints, and feed into the LSTM. The feature pooling techniques have been applied for unimodal images based on the consistent contents, such as temporal pooling of frame-wise features in video recognition [18] and multi-view pooling in 3D shape recognition [19] whereas we focus on the multi-perspective images. It can be easily applied various existing methods and pre-trained models as is, however, some unique attributes in each image may be diminished by pooling.

Consequently, we propose uni-supervised and tri-regularized training which aims to train the multi-perspective model without fully-paired supervision (See Fig. 3). It consists of the uni-supervision stream that leverages general image-caption pairs and the tri-regularization stream that self-supervises by the triplet images. On the uni-supervision stream, we first make a masked feature instead of the f in Eq. 2 by randomly select a view as an image feature and setting other views to zero ("Switch View" in Fig. 3). Furthermore, all the nodes in the selected view are scaled by factor 3. Note that we change this multiplexer mechanism to an identity function during testing. Next, we propagate through the remained networks and computes the crossentropy loss $\mathcal{L}_{CE}$ indicated in Eq. 5. Whereas, the tri-regularization stream does not use the image-caption dataset but sets of triplet images captured from the actual scenes. It configures the autoencoder, containing the fusion layer as an encoding part. This stream aims to preserve the information in three views and avoid simple averaging in the uni-supervision stream. Thus we compute the reconstruction loss $\mathcal{L}_{MSE}$ defined by mean square error. From the CNNs to the fusion layer, both streams share the parameters to be learned. The objective is to minimize a weighted loss of $\mathcal{L}_{CE}$ and $\mathcal{L}_{MSE}$.

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{MSE} \tag{6}$$

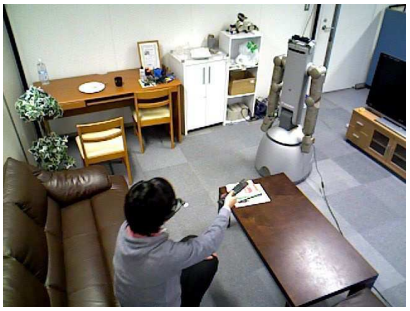


Fig. 4: Experiment room we use for data collection

V. Experiments

A. Data

We use MSCOCO dataset [20] to train our model on the uni-supervision stream. We use the data splits from [21] which contains 113,287 images for training, 5,000 images for validation, and 5,000 images for testing. Each image has 5 captions. As a preprocessing of the sentences, we remove punctuation and unify them in lower case. We prepare a `<start>` and a `<end>` tokens as special words and prune the vocabulary by defining any words that have a count less than 5 as a special `<unknown>` word. The final vocabulary comprises 10,107 words.

Furthermore, to train our model on the tri-regularization stream, we collected multi-perspective image sets in the experiment room shown in Fig. 4. The room is filled with various items of furniture such as a desk, a couch, a plant, and a television. We assumed that there are one resident and one service robot in the room. To assemble and collect the triplet images, a person wore a glasses-type wearable camera (Vuzix M100 [22]) for streaming the first-person images to a cloud server and acted some activities of daily living such as reading a book, walking around, and watching television. Moreover, we collected the second- and third-person images, from a camera in the head of the service robot (Yaskawa SmartPalV [23]) and cameras fixed on the wall, respectively. Each image is 360×480 resolution. We simultaneously recorded three type of images in every 5 seconds, with 8 different camera positions. The final image sets contain 200 triplets for training and extra triplets in several situations for testing. Example images are shown in Fig. 1.

B. Implementation Details

The parameters of image encoder are transferred from ResNet-152 [24] pre-trained on ImageNet [16] dataset up to the global average pooling layer. Thus, given an arbitrarily sized color image, the outputted image feature is a 2048-dimensional vector. The fusion layer receives a concatenated 6144-dim vector from the image encoder and projects it to 2048-dim feature space (see Eq. 1). The input and hidden state dimensions of the LSTM are set to be 512.

We compute the weighted loss from the crossentropy of MSCOCO data and the self-supervised mean square error of the triplet images, each of which is with a mini-batch of 16. We choose the weight of 1 for λ of Eq.6. We train our model under the weighted loss objective using Adam optimizer with a learning rate of 5×10^{-4}. We decay the learning rate by multiplying factor 0.8 in every 3 epochs. Moreover, we employ "Scheduled Sampling" [25] where the model feeds back its own prediction to input at next step under the feedback probability p. We initialize the factor p with 0 and increase by 0.05 every 5 epochs until it reaches 0.25. We select the final model which achieves the best CIDEr-D score on *duplicated-input* images made of MSCOCO validation set.

All algorithms are implemented in PyTorch [26] and run on NVIDIA GeForce GTX Titan X.

TABLE I: Caption scores on MSCOCO test splits.

Input	BLEU-4	METEOR	ROUGE-L	CIDEr-D	SPICE
Single	27.8	24.3	51.8	90.9	17.6
Triplet†	26.0	23.0	50.3	83.2	16.2

† Set an identical image to three inputs.

C. Results

We compare our multi-perspective model with the basic single-perspective model which receives only one image. We herein denote the models as "Triplet" and "Single", respectively. In Table I, we show some qualitative scores on the MSCOCO test split with common metrics in natural language processing tasks. We generated captions by greedy decoding. Note that our scores here are not in the multi-perspective scenario. It can be seen that the performance gap is small even with the random view switching and the regularization during training our model.

In addition to caption generation, we visualize the internal response in models by Grad-CAM [27]. Grad-CAM enables us to see discriminative regions that strongly influence a resulted caption, by weighting feature maps of a certain layer with gradients derived from the score. We choose the last convolution layer of the image encoder for extracting feature maps. Moreover, we extend the Grad-CAM calculation by computing word-wise log probability for each step instead of sentence-wise log probability [27]. We normalize the gradient-based weights over three streams of all the steps.

We generated captions from our triplet images with (i) the single-perspective model which produces for each image (Single), (ii) the same model with a pooled feature (Single-Pool), and (iii) our proposed multi-perspective model (Triplet). We used beam search approach to decode words with a beam width of 5. The example of generated captions and the Grad-CAM visualization is shown in Fig. 5. In the situation of Fig. 5, the man sitting on the couch is reading the book, otherwise, the robot is facing to the man and they are in the living room. (i) As shown in Fig. 5(a)(b)(c), given the first-, second-, and third-person images, the single-perspective model produced *"a person holding a pair of scissors in their hand"*, *"a woman sitting on a couch with a cat in her lap"*, and *"a man sitting in a chair in a living room"*, respectively. Although each caption deviates from the actual scenes, they captured fragmentary

visual concepts, for instance, the fact that the actor holding a object, which has appeared in the first-person vision. (ii) As shown in Fig. 5(d), the feature-pooling method produced *"a woman sitting on the floor playing a video game"*. It succeeded to describe the sitting man but the remained sentence is totally incorrect. According to the visualization, the phrase *"playing a video game"* was affected by the region around the robot. (iii) Finally, as shown in Fig. 5(e), our proposed multi-perspective model produced *"a woman sitting on a couch using a laptop computer"*. The result was still inappropriate but succeeded to describe the woman sitting on a couch from the triplet images with only the single supervision, and to strongly and simultaneously activate in the couch regions on the second- and third-person images. Additionally, we provide other results with four types of models in Fig. 6.

VI. Conclusion and Discussion

In this paper, we have proposed a novel image captioning approach using a multi-perspective viewpoint, *i.e.*, the first-person image from a wearable camera, the second-person image from a camera mounted on a robot, and the third-person image from an embedded camera in the intelligent space. To generate captions from the triplet images, we extended the basic encoder-decoder network and introduced uni-supervised and tri-regularized training which enables us to train the model with only single-perspective annotations. Our model succeeded to learn to generate caption from the triplet images under the training strategy we proposed. Although we could confirm some improvement in some samples, even the results from single images contained inappropriate words. It can be considered that the image quality or domain gap between datasets caused the problems. Future work includes constructing an annotated test set for evaluation and demonstrate the effectiveness quantitatively.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number JP26249029 and JST CREST Grant Number JPMJCR17A5, Japan.

References

[1] J.-H. Lee and H. Hashimoto, "Intelligent space—concept and contents," *Advanced Robotics*, vol. 16, no. 3, pp. 265–280, 2002.

[2] M. Niitsuma and H. Hashimoto, "Spatial memory as an aid system for human activity in intelligent space," *IEEE Transactions on Industrial Electronics*, vol. 54, no. 2, pp. 1122–1131, 2007.

[3] R. Kurazume, Y. Pyo, K. Nakashima, A. Kawamura, and T. Tsuji, "Feasibility study of iort platform" big sensor box" ," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 3664–3671.

[4] M. Bolanos, M. Dimiccoli, and P. Radeva, "Toward storytelling from visual lifelogging: An overview," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 1, pp. 77–90, 2017.

[5] T. Miyanishi, J.-i. Hirayama, Q. Kong, T. Maekawa, H. Moriya, and T. Suyama, "Egocentric video search via physical interactions," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16. AAAI Press, 2016, pp. 330–336. [Online]. Available: http://dl.acm.org/citation.cfm?id=3015812.3015861

[6] H. Nakayama, T. Harada, and Y. Kuniyoshi, "Ai goggles: Real-time description and retrieval in the real world with online learning," in *Canadian Conference on Computer and Robot Vision (CRV)*. IEEE, 2009, pp. 184–191.

[7] C. Fan and D. Crandall, "Deepdiary: Automatically captioning lifelogging image streams," in *European Conference on Computer Vision International Workshop on Egocentric Perception, Interaction, and Computing (EPIC)*, 2016.

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 652–663, April 2017.

[9] M. S. Ryoo and L. Matthies, "First-person activity recognition: What are they doing to me?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 2730–2737.

[10] R. Yonetani, K. M. Kitani, and Y. Sato, "Recognizing micro-actions and reactions from paired egocentric videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2629–2638.

[11] K. Nakashima, Y. Iwashita, P. Yoonseok, A. Takamine, and R. Kurazume, "Fourth-person sensing for a service robot," in *Proceedings of the IEEE Conference on Sensors*, 2015, pp. 1–4.

[12] C. Fan, J. Lee, M. Xu, K. Kumar Singh, Y. Jae Lee, D. J. Crandall, and M. S. Ryoo, "Identifying first-person camera wearers in third-person videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[13] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 3209–3216.

[14] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, and L. Fei-Fei, "Detecting events and key actors in multi-person videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3043–3053.

[15] M. Bolaños, Álvaro Peris, F. Casacuberta, S. Soler, and P. Radeva, "Egocentric video description based on temporally-linked sequences," *Journal of Visual Communication and Image Representation*, vol. 50, pp. 205–216, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047320317302316

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, USA, June 2009, pp. 248–255.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2015, pp. 448–456.

[18] M. S. Ryoo, B. Rothrock, and L. Matthies, "Pooled motion features for first-person videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 896–904.

[19] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 945–953.

[20] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[21] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3128–3137.

[22] "Vusix M100," https://www.vuzix.com/Products/m100-smart-glasses.

[23] "Yaskawa SmartPalV," https://www.yaskawa.co.jp/en/.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

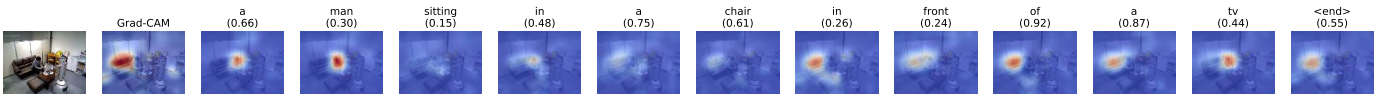
[25] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 1171–1179.

[26] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

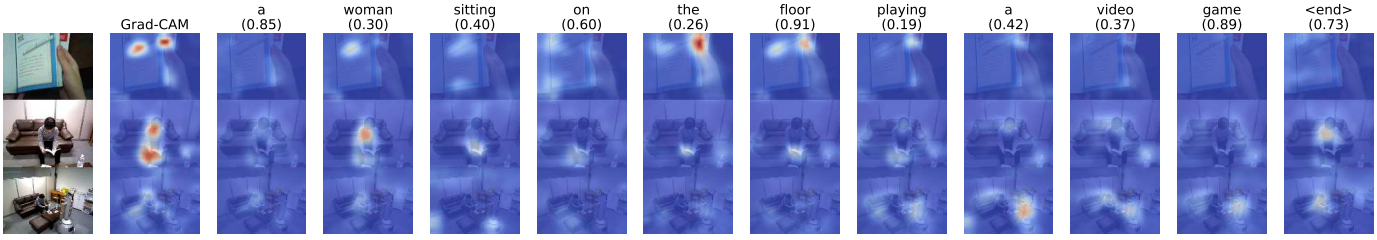
[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

(a) The first-person viewpoint by the model *Single*

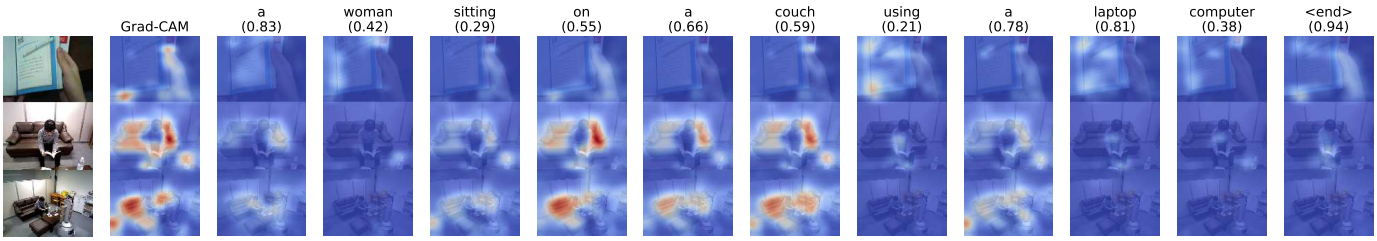
(b) The second-person viewpoint by the model *Single*



(c) The third-person viewpoint by the model *Single*



(d) The fourth-person viewpoint by the model *Single-Pool*



(e) The fourth-person viewpoint by the model *Triplet* (**ours**)

Fig. 5: Results from triplet images captured in the intelligent room. The first and the second figures from the left are input images and sentence-wise Grad-CAM [27]. The following figures from the third are word-wise Grad-CAM, each of which is with the predicted word and its probability. Regions highly contributing to predict the word appears in **red** color.

Example 1

Single-1: a close up of a person holding a skate board, **Single-2:** a couple of people that are sitting on **a couch**, **Single-3:** a group of people sitting **around a table in a room**, **Single-Pool:** a woman sitting in front of a laptop computer, **Triplet:** **a man** sitting on **a couch in a living room**

Example 2

Single-1: a close up of a black and white object, **Single-2:** **a woman sitting at a desk** with a laptop, **Single-3:** **a living room with furniture and a flat screen tv**, **Single-Pool:** **a person sitting at a desk** with a laptop, **Triplet:** **a desk** with a computer and a lamp on it

Example 3

Single-1: **a person holding** a nintendo wii game controller, **Single-2:** **a woman sitting at** a table with a cup of coffee, **Single-3:** **a living room filled with furniture and a flat screen tv**, **Single-Pool:** **a man sitting at** a table in front of a laptop computer, **Triplet:** **a man sitting at a desk** in front of a computer

Example 4

Single-1: **a close up of a person holding** a cat, **Single-2:** **a woman** standing in front of a refrigerator, **Single-3:** **a living room filled with furniture** and a fire place, **Single-Pool:** **a person sitting on** a chair in front of a tv, **Triplet:** **a man** standing in a kitchen next to a cat

Fig. 6: Generated examples, each of which is with input images (left) and outputted captions (right). The each image is a set of the first- (top), second- (middle), and third-person images (bottom). The "Single-X" denotes a caption from the Xth-person image. True positive terms are highlighted in **blue**. Some improvements by Triplet model (ours) can be seen in the examples 1 and 3. Whereas the examples 2 and 4 fail to generate appropriate captions.