Self-supervised Human–Object Interaction of Complex Scenes with Context-aware Mixing: Towards In-store Consumer Behavior Analysis

Takashi Kikuchi, Shun Takeuchi
Fujitsu Research
4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki-shi, Kanagawa, Japan
{kikuchi_takashi, shun.takeuchi}@fujitsu.com

Abstract

Recognizing human–object interactions (HOIs) in physical retail stores, such as picking up a product, can provide valuable information about non-purchasers, and is an important aspect of understanding customer behaviors. However, there are often complex scenes in physical retail stores with numerous similar objects in the shelf, making the task of recognizing the interacting object challenging. To address the drawback of complex background scenes, we propose a method using image mixing and self-supervised techniques to train the model to differentiate objects that interact with background objects. The proposed method generates images without the object's influence based on the input image using Context-aware image mixing. Then, we introduce a self-supervised method using the generated images to learn the difference between the actual and the background objects. We evaluated the network's performance using public and private retail dataset. We confirmed that when applied to physical retail scenes, the performance overcame the recent HOI detection methods including the recent state-of-the-art method. To the best of our knowledge, this is the first study to apply a self-supervised technique to control the target of interaction for the HOI detection model, demonstrating promising potential for use in in-store consumer behavior analysis.

1. Introduction

In recent years, understanding customer behavior in physical retail stores has been recognized as an important aspect of promoting sales for the retailers [15]. Numerous studies have been conducted to utilize Point of Sale (POS) data in devising sales plans [14,21]. However, POS data primarily contains information about purchasers, while information about non-purchasers is lost. Consequently, research on customer behavior recognition using video data has been carried out to understand the purchasing tendencies of non-purchasers [1, 16]. Among these, information about which products they reached for is essential for discerning customer interests, necessitating the recognition of interactions between people and objects.

Figure 1. (a) Primary images found in the public dataset, where target objects are positioned at the center. (b) Typical failure cases of conventional methods. Complex scenes with similar objects make the HOI detection task complicated.
Left: Predicted image, Right: Ground-truth image.

Human–object interaction (HOI) detection has recently attracted attention, and is considered for integration into various applications due to its high potential to deeply understand image scenes. Given an image, HOI detection predicts a set of <human, object, action> HOI triplet, which is a task to localize a human and object, and extract the semantic relationship between them.

However, when considering its practical deployment in real-world scenarios, the intricacies of the background complexity exert a substantial influence on the

precision of these methods as shown in Figure 1. For instance, in domains of retail scenes, the contextual surroundings pose heightened challenges to accurate detection [12]. The first challenge arises from the proliferation of objects similar to those actively involved in the scene, which complicates the demarcation between foreground entities and the background. This convolution in distinguishing the object of interest stems from the abundance of akin objects proximate to the target. For example, when discerning instances of interactions involving products like wine lifted from a merchandise shelf, the presence of diverse commodities in close proximity confounds the accurate determination of the specific item manipulated. The second challenge revolves around the density of objects within the background. Even when no analogous objects are in the immediate vicinity, the presence of shelves laden with merchandise contributes to the occlusion of objects, inducing false-positive identifications across various ranges in the surroundings.

Hence, providing a dataset specific to the scene becomes imperative to instill the differentiation between the background and the interacting objects to address scenarios of this nature. However, creating datasets for HOI detection proves to be an intricate and time-consuming endeavor due to the multifaceted nature of the tasks involved. Furthermore, a substantial quantity and diversity of training data are requisite for effective learning of transformers. Moreover, the characteristics of the background are contingent upon factors such as the camera's field of view and the setting, thereby necessitating the generation of datasets on a per-scene basis to uphold precision.

In this paper, we propose a HOI detection method that leverages existing publicly available datasets to maintain a certain level of accuracy, even in scenes featuring intricate backgrounds. The key tenet of our proposed methodology lies in integrating of self-supervised learning, wherein the disparities between feature representations when objects are present in the vicinity and when they are not are learned. Our method generates multiple synthetic images using an image mixing method, called "Context-aware mixing", wherein the same object near an interacting object is composited into the input image. By incorporating self-supervised learning, it learns to distinguish scenarios with and without analogous background objects.

This approach enables the differentiation and learning of features associated with the presence of similar objects in the surrounding environment, a notable characteristic in scenarios like retail scenes. Thus, our proposed method aims to enhance accuracy in complex backgrounds replete with similar objects, a challenge prevalent in the aforementioned retail domain.

This paper's contributions are summarized as follows:

- We propose a transformer-based HOI method with self-supervised learning to differentiate the target and background objects.
- We introduce an image mixing method (Context-aware mixing) to generate challenging images with non-interactive objects nearby that lead to potential confusion.

2. Related Work

2.1. HOI Detection

HOI detection task can be defined as the combination of object detection and interaction recognition. Based on this definition, the existing HOI detection methods can be summarized into two approaches: two-stage approaches and one-stage approaches.

Two-stage HOI detection approach allows the detection task to divide into two individual tasks: object detection and interaction recognition. The object detection task involves localizing targets using a generic object detector, preparing human-object pairs. Then, classification of interaction labels are conducted for each pair by cropping features of the backbone network within the localized region. Furthermore, additional features such as human pose [9, 26] are used as supplementary information and graph-based methods [20, 24] are also used to understand complex relationships between humans and objects. However, due to the large number of human-object pairs handled, two-stage approach encounter the problem of expensive computation.

One-stage HOI approach detects directly all the HOI triplets by performing object detection and interaction prediction in parallel. Recently, Transformer-based HOI detectors has become the main approach for HOI detection due to the appearance of DETR [2]. Tamura et al. [19] proposed a method based on DETR to use self-attention mechanism of the transformer to extract contextual semantic information and the embeddings to predict HOI instance. However, despite the abundance of transformer-based methods, there remains a noticeable gap in the literature when it comes to addressing the challenge of background objects that are often occluded by the dataset content. This particular aspect, where background objects play a crucial role, has yet to receive significant attention within the existing body of research.

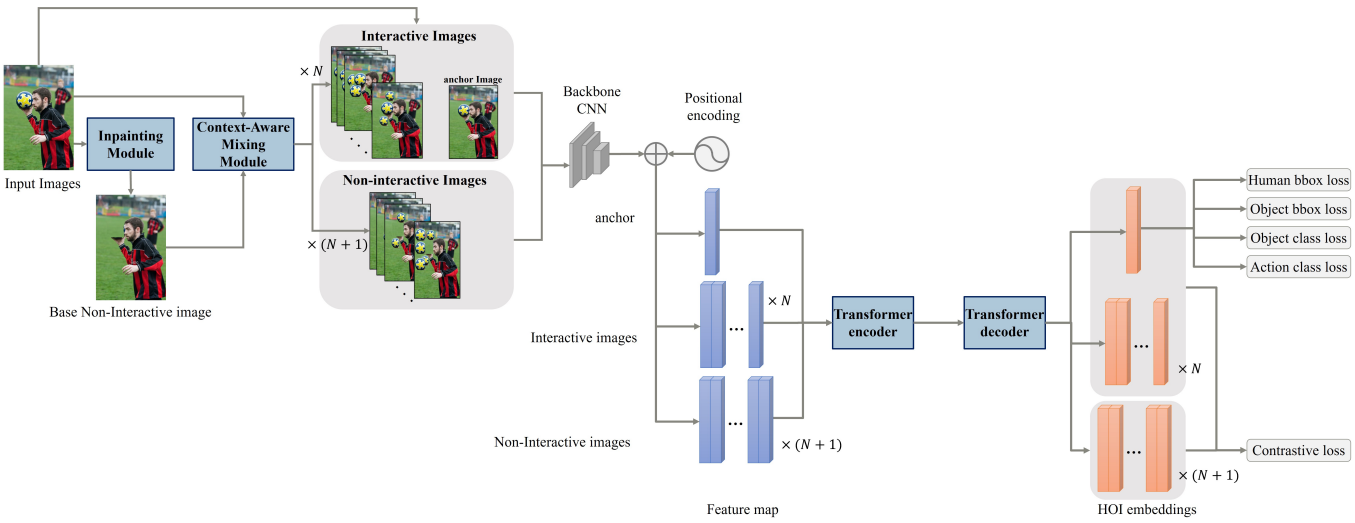

Figure 2. Overall architecture of the proposed method. The Context-Aware Mixing module generates challenging images with non-interactive objects nearby. Using the set of images with interacting objects and images without interacting objects, the model is trained to learn to differentiate the target object with surrounding objects by supervised contrastive learning.

2.2. Self-supervised Approach with HOI Detection

The motivation for self-supervised learning is to reduce human labeling and achieve performance comparable to supervised learning without the need for explicit labels. Among various approaches, contrastive learning is a method that learns image features without using labels. It ensures that similar data have similar embeddings and different data have different embeddings. This method is utilized in various scenarios [18].

Contrastive learning techniques are also employed in the HOI detection field. Zong et al. [27] used contrastive learning to reduce misclassification between unseen objects and confusable seen objects. However, in their approach, contrastive learning was used to bridge the gap between seen and unseen objects and to solve zero-shot learning. In contrast, our method leverages contrastive learning to recognize HOIs with targets in complex backgrounds.

3. Method

3.1. Overview

The architecture of the proposed method is illustrated in Figure 2, which is based on the transformer-based model, where the input image is augmented and trained with contrastive learning. Given an input image, several images are processed by context-aware image mixing by utilizing the processed images. The model is trained to learn to differentiate objects that interact with background objects. We adopt QPIC [19] as the base model with ResNet-50 [5] as the backbone of the CNN. In addition, contrastive loss is adopted to the base model as the loss function.

3.2. Context-aware Mixing

In this subsection, we introduce the image augmentation method for training the difference between the interactive and background objects. In the image augmentation process, two sets of images are generated: human–object interactive and non-interactive image. The interactive image is the typical HOI input image, where the person interacts with an object. However, non-interactive image has objects around but the person is not interacting with any of them.

An interactive image is generated based on the input image. This image augmentation aims to create images resembling complex scenes where multiple similar objects surround the target object. Typically, image mixing involves randomly patching in portions from other images [7, 25] or processing within the same image. However, we aim to produce images resembling intricate scenes, making entirely random processes inadequate for our requirements. Inspired by SalfMix [4], our approach, context-aware mixing introduces a self-guided technique for image mixing. The key distinction from SalfMix lies in our requirement, which is object-centric, utilizing object positions derived from the annotation labels. Furthermore, we impose constraints on where the objects are patched: these positions are proximate to the target object. This tailored augmentation approach enhances the contextual accuracy of

Figure 3. (a) Example images of context-aware mixing. The object area is cropped and is mixed nearby. This process is conducted only in training phase. (b) Example images of non-interactive images. (c) Example of the result when applying context-aware mixing data to pretrained HOI detection model (Left: Prediction Image, Right: Ground-truth image). The original object is diminished by the inpainting method.

our generated images.

In the actual synthesis process, we perform random resizing and rotation of the original object at a distance of Δd from its initial position. Before synthesis, a fixed-size zero-padding is applied around the image, and after synthesis, the padded regions are removed, allowing the synthesized portions to exist within the image boundaries. We set two hyperparameters to implement this: $\max \Delta d$ (maximum distance) and n (number of objects). The distance Δd is selected randomly within the range of $[0, \max \Delta d]$.

Non-interactive images are also generated based on the input images. Our objective is to learn subtle distinctions between the target and background objects. To achieve this, we create necessary images by masking the target object from the original image and extending the surrounding identical objects using context-aware mixing. Given that objects manipulated by hand frequently involve humans, especially to mitigate the influence of actions, we utilized inpainting techniques [22] to delicately obscure only the target object. We employed this image as the base image for non-interactive images and as the input image for context-aware mixing. This approach ensures that the human-associated regions are preserved as much as possible, while the masked regions undergo smoothing. Compared to simple cropping or blurring, this inpainting procedure enables a more natural extension of the image, thus yielding a refined augmentation methodology.

The examples of the Context-Aware Mixing are shown in Figure 3. In the presence of multiple objects augmented near the original object, it has been observed that the prediction results exhibit variations when employing context-aware mixing. As depicted in Figure 3(c), the prediction outcome is influenced by the presence of nearby objects. To address this issue, we utilize these instances to train the model to effectively distinguish between the target and background objects. This phenomenon highlights the potential influence of surrounding objects on the model's performance and warrants further investigation to enhance the robustness of the approach.

3.3. Self-supervised Learning

Our method involves concurrently implementing self-supervised learning, specifically contrastive learning, alongside the conventional training of a transformer-based HOI model. In contrastive learning, for each input image, we learn the disparities between images generated using the approach elucidated in the preceding subsection 3.2—images interacting with objects and images without interactions. In the context of contrastive learning, we focus on enhancing the distinction between two sets of images: those interacting with objects and those without interactions. This process enables the model to learn features that can differentiate instances where interactions occur from instances where they do not.

In our approach, within a single input image, we engage in a two-class contrastive learning process, wherein the learning pertains to the differences between images depicting interactions and those without of interactions. In this context, the two classes pertain to images illustrating actions being executed and images in which no actions are occurring. We adapt the method of Khosla et al. [6] to accomodate this,which uses supervised contrastive learning, a contrastive learning technique capable of treating multiple images as belonging to the same class.

Supervised contrastive loss is added to the original HOI losses and is expressed as in equation (1),

$$L_{\text{sup}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)}. \tag{1}$$

here, L_{sup} represents the loss of supervised contrastive

Content of the Dataset	mAP HICO-DET
Base Dataset	58.6
Add inpainting image (+1 image)	56.2
Add Context-aware mixing images (+5 images)	43.2

Table 1. Results of the base model when simply adding context-aware mixing Data. Used dataset that only contains the "hold" action.

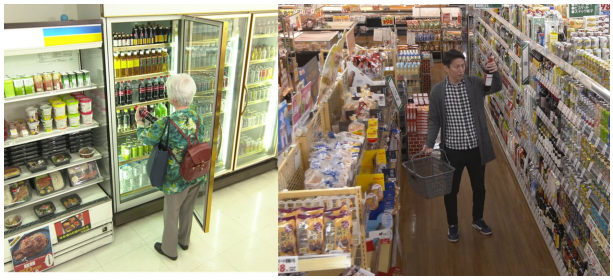

Figure 4. Example images of the private dataset. The dataset consists of "holding product" scenes from the CCTV camera of real retail store.

learning, i denotes the anchor index, A refers to I/i, which includes all images except for the anchor image, $P(i)$ is the set of indices of all positive samples, i.e., all images sharing the same label as the anchor, z is the embeddings of the HOI instance and τ is the scalar temperature parameter.

By integrating this contrastive learning scheme, we enrich the discriminative capacity of our approach and further elevate its efficacy in capturing intricate contextual relationships.

3.4. Loss Calculation

Our proposed method incorporates the losses from the base model (QPIC) [19] and augments them with an additional contrastive loss (L_{sup}) explained in subsection 3.3. For the bounding box predictions in human and object recognition tasks, the L1 loss (L_{b}) and Generalized Intersection over Union (GIoU) [17] loss (L_{u}) are employed. The cross-entropy error (L_{c}) is utilized for object category classification, while the focal loss (L_{a}) is adopted for action classification. Following the procedure of QPIC, the Hungarian algorithm [8] is used for bipartite matching.

Transformers necessitate large datasets for effective training and are sensitive to the quality of the data provided. As illustrated in Table 1, the prediction performance deteriorates when merely incorporating augmented images into the dataset. Unlike conventional data augmentation techniques, we hypothesize that this phenomenon is attributed to overfitting, as the augmented images do not introduce significant variations compared to the original images. This result also demonstrates that simply adding complex images to a dataset is insufficient to improve performance when the complex data is added randomly. It is necessary to add data that is relevant to the domain to effectively improve performance. Consequently, the prediction performance declines as the number of images in the dataset increases, emphasizing the need for more diverse and representative data samples to mitigate overfitting and improve generalization. To address this issue, we computed the loss for the recognition and classification tasks (L_{b}, L_{u}, L_{c}, L_{a}) using only the original input images. Meanwhile, the images generated by context-aware mixing were exclusively employed for the calculation of the contrastive loss (L_{sup}).

4. Experiments

4.1. Datasets and Evaluation Metrics

Our experiments employed the commonly used HICO-DET [3] dataset as the training dataset for HOI validation. We utilized two distinct datasets for evaluation: the public dataset, HICO-DET, and a private dataset featuring scenes with complex backgrounds, sourced from retail stores. The public dataset, like HICO-DET, encompasses scenes with intricate backgrounds, but a significant portion of the data primarily showcases interactions between people and target objects at the center of the frame.

However, the complex scenes where our method is focused on discerning target and background objects are relatively scarce. Therefore we introduced the private dataset to adequately assess accuracy in such scenarios. This private dataset emulates physical retail store environments with intricate backgrounds, drawn from actual footage of retail stores. For the physical retail store dataset, we created two datasets. The first dataset comprises imagery from multiple stores, each distinct in layout despite sharing the same store identity. The second dataset was generated within the context of the first dataset, specifically focusing on creating images with complex scenarios involving store layouts and camera angles. For each dataset under consideration, a total of 100 images were collected. The images for both dataset are collected from the CCTV cameras of the store and were annotated manually.

Method	Backbone	HICO-DET	Private Retail Dataset	
			Multiple retail stores	Complex retail shelf
QPIC [19]	ResNet-50	58.6	34.8	28.1
CDN [23]	ResNet-50	**61.2**	<u>36.6</u>	30.7
GEN-VLTK [10]	ResNet-50	<u>60.5</u>	35.8	<u>37.5</u>
Our method	ResNet-50	53.3	**47.8**	**39.2**

Table 2. Comparison with the recent methods.
The top mAP is shown in bold and the second with underline

Given that the private dataset predominantly revolves around the action of "holding"—specifically, the act of picking up products—we decided to center the entire evaluation process around this "hold" action. Moreover, due to constraining the action class to a single category, there might not be sufficient images available for training all object categories for the transformer-based model. To mitigate this limitation, we have unified all object categories as "Something". This approach primarily emphasizes the discernment of objects related to the action, while maintaining the object categories at a par with conventional methods. Consequently, in this experiment, we believe that the impact of defining all object categories as "Something" does not significantly affect the results, as the core focus lies on the determination of objects involved in the action.

In evaluating HOI instances, we adopted the commonly used mean average precision (mAP) metric prevalent in existing HOI research. Like other HOI researches, the HOI instance which is considered to be the true positive (TP) has all the following condition satisfied.

- The category labels of human and object pairs are both correct
- The intersection over union (IoU) between the ground-truth annotation exceed 0.5 for both human and object, i.e. $\min(\mathrm{IoU}_{\mathrm{human}}, \mathrm{IoU}_{\mathrm{object}}) > 0.5$
- The interaction label is correct

Moreover, due to the unification of object categories as "Something", we refrained from utilizing frequency-based categorization as employed in other studies.

4.2. Implementation

We used QPIC [19] as a base model for our method and used ResNet-50 as a backbone feature extractor. The encoder and decoder for the transformer have six layers, and the number of query embeddings is 100. Before the training, we initialized the network with a parameter of DETR pretrained with the COCO dataset [11]. We trained the model for a total of 150 epochs employing a batch size of 16 (two images per GPU with 8 GPU). We used the AdamW [13] optimizer with the backbone's learning rate of 10^{-5} and the other's 10^{-4} and the weight decay 10^{-4}. The hyper-parameters of the adjusting loss weights λ_{b}, λ_{u}, λ_{c}, λ_{a} λ_{sup} are set to 2.5, 1, 1, 1, 1 and the hyper-paramter for Context-Aware Mixing max Δd is set to 100. The learning rates for both the backbone and the others are decayed after the 100th epoch.

Method	Number of mixing object	HICO-DET	Multiple retail stores
Base-model	0	58.6	34.8
Our method	1	49.0	28.8
	5	53.3	30.2
	10	53.3	**47.8**
	15	**57.2**	35.8
	20	56.8	37.8

Table 3. Comparison with the number of objects mixed to the image.

4.3. Comparison with the recent methods

We compared our method with the recent HOI methods as shown in Table 2. Compared with the other transformer-based one-stage method, our model showed slightly lower results for the public dataset HICO-DET. However, when applied to complex scenes our result exceeded the other method by 12.0 mAP (relatively 33.5%) for multiple retail scenes and 1.7 mAP (relatively 4.5%) compared with GEN-VLTK [10], the recent state-of-the-art method. Also, when comparing the result with the base model (QPIC) our method has exceeded by 13.0 mAP (relatively 37.4%) for multiple retail scenes and 11.1 mAP (relatively 39.5%) for the complicated scene.

From the results, it can be inferred that our method enabled the model to learn more effectively. An example of our method's output is depicted in Figure 5. Evidently, our approach successfully prevented the model from predicting background objects as interactive objects. However, this had an adverse effect on the pub-

Figure 5. Examples of the results. (a):Prediction results of the base model [19]. (b): Prediction results of our method. Green bounding box: human, Blue bounding box: object.

lic dataset case, where some items were not recognized and were instead predicted as part of the background. One factor leading to the reduced accuracy is that some objects were recognized as background entities, resulting in their non-detection. This highlights the need to further refine our method to address such issues and improve object detection performance, particularly in complex scenes. Certainly, in HICO-DET, which comprises numerous images with the target person or object centrally positioned, there is a decrease in performance. Nonetheless, a substantial improvement in performance is evident for complex scenes, indicating an overall enhancement in generalization ability. In this paper, we focused on the action "hold" and object "something" due to the training data, so we will expand our method to detect more actions and objects for the future work.

We also conducted an experiment to investigate the impact of varying the hyperparameter controlling the number of objects mixed with each image on the precision of our object detection model. As illustrated in Table 3, the precision of the model decreases when the number of objects is insufficient. This finding emphasizes the importance of carefully selecting the appropriate number of objects for training deep learning models in object detection tasks. We hypothesize that the decrease in precision when the number of objects is insufficient may be due to the transformer's inability to fully understand the local differences in the images, as the variations are relatively small. This results in the model being negatively affected by the input of low-quality images. Furthermore, when a certain number of objects are synthesized, the precision improves compared to the original model. However, we interpret that the excessive density of objects may make it difficult to discern the target objects, leading to a decrease in precision. This suggests that there may be an optimal balance between the number of objects and the model's performance in object detection tasks.

5. Conclusion

In this paper, we have proposed a transformer-based HOI approach with self-supervised learning to differentiate target and background objects. The proposed method introduce a image mixing method, context-aware mixing, to generate confusing image that looks as if the human in the image is interacting to other object in the background. Through our experiments, we verified that our methodology outperformed other techniques including the recent state-of-the-art method, particularly excelling in complex scenes as those found in retail stores. We confirmed that the generalization performance has been improved by the proposed method. Furthermore, our methodology is designed to complement existing transformer-based approaches, allowing for additional learning enhancements. Our approach can be applied beyond the specific method we've presented here, potentially enhancing various transformer-based models in different contexts. Additionally, Our methodology does not impact the inference flow of the base model, and as a result, the processing speed remains comparable to that of the base

model. Therefore, the introduction of our approach holds the potential for real-time applications, as it does not compromise the processing efficiency of the underlying model. In physical retail stores, recognizing and analyzing consumer behavior can lead to the provision of desired products for the shoppers. We believe that this technology holds great potential for being utilized effectively in such contexts.

References

[1] Almustafa Abed, Belhassen Akrout, and Ikram Amous. Shoppers interaction classification based on an improved densenet model using rgb-d data. In 2022 8th International Conference on Systems and Informatics (ICSAI), pages 1–6, 2022. 1

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. 2

[3] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions, 2018. 5

[4] Jaehyeop Choi, Chaehyeon Lee, Donggyu Lee, and Heechul Jung. Salfmix: A novel single image-based data augmentation technique using a saliency map. Sensors, 21(24), 2021. 3

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 3

[6] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021. 4

[7] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In Hal Daumé III and Aarti Singh, editors, Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 5275–5285. PMLR, 13–18 Jul 2020. 3

[8] Harold W. Kuhn. The Hungarian Method for the Assignment Problem. Naval Research Logistics Quarterly, 2(1–2):83–97, March 1955. 5

[9] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Xijie Huang, Liang Xu, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1–1, 2021. 2

[10] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection, 2022. 6

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 6

[12] Jun Liu, Ye Liu, Guyue Zhang, Peiru Zhu, and Yan Qiu Chen. Detecting and tracking people in real time with rgb-d camera. Pattern Recognition Letters, 53:16–23, 2015. 2

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 6

[14] Thomas Edward Muller. Structural information factors which stimulate the use of nutrition information: A field experiment. Journal of Marketing Research, 22:143 – 157, 1985. 1

[15] Marina Paolanti, Daniele Liciotti, R. Pietrini, Adriano Mancini, and Emanuele Frontoni. Modelling and forecasting customer navigation in intelligent retail environments. Journal of Intelligent & Robotic Systems, page 1–16, 10 2017. 1

[16] Marina Paolanti, Rocco Pietrini, Adriano Mancini, E. Frontoni, and Primo Zingaretti. Deep understanding of shopper behaviours and interactions using rgb-d vision. Machine Vision and Applications, 31, 2020. 1

[17] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression, 2019. 5

[18] Madeline C. Schiappa, Yogesh S. Rawat, and Mubarak Shah. Self-supervised learning for videos: A survey. ACM Computing Surveys, 55(13s):1–37, jul 2023. 3

[19] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information, 2021. 2, 3, 5, 6, 7

[20] Oytun Ulutan, A S M Iftekhar, and B. S. Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions, 2020. 2

[21] Arch Woodside and G.L. Waddle. Sales effects of in-store advertising. Journal of Advertising Research, 15:29–33, 01 1975. 1

[22] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting, 2023. 4

[23] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection, 2021. 6

[24] Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions, 2021. 2

[25] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018. 3

[26] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. Polysemy deciphering network for robust human-object interaction detection, 2021. 2

[27] Daoming Zong and Shiliang Sun. Zero-shot human–object interaction detection via similarity propagation. IEEE Transactions on Neural Networks and Learning Systems, pages 1–12, 2023. 3