WHU-NERCMS AT TRECVID2021: INSTANCE SEARCH TASK

Yanrui Niu[†], **Jingyao Yang**[†], **Ankang Lu**[†], **Baojin Huang**[†],
Yue Zhang, **Ji Huang**, **Shishi Wen**, **Dongshu Xu**, **Chao Liang**[*], **Zhongyuan Wang**[*], **Jun Chen**[*]
Hubei Key Laboratory of Multimedia and Network Communication Engineering
National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
cliang@whu.edu.cn

ABSTRACT

We will make a brief introduction of the experimental methods and results of the WHU-NERCMS in the TRECVID2021 in the paper. This year we participate in the automatic and interactive tasks of Instance Search (INS). For the automatic task, the retrieval target is divided into two parts, person retrieval, and action retrieval. We adopt a two-stage method including face detection and face recognition for person retrieval and two kinds of action detection methods consisting of three frame-based human-object interaction detection methods and two video-based general action detection methods for action retrieval. After that, the person retrieval results and action retrieval results are fused to initialize the result ranking lists. In addition, we make attempts to use complementary methods to further improve search performance. For interactive tasks, we test two different interaction strategies on the fusion results. We submit 4 runs for automatic and interactive tasks respectively. The introduction of each run is shown in Table 1. The official evaluations show that the proposed strategies rank 1^{st} in both automatic and interactive tracks.

Table 1: Result of each run

Type	Run ID	Relation	mAP	Strategy
Automatic	F_2	–	**0.435**	A + F + S + R
	F_6	–	0.418	A + F + S
	F_4	–	0.418	A + F + S (w/o STE on kissing)
	F_8	–	0.395	A + F
Interactive	I_1	F_2 + Top-K	**0.465**	A + F + S + R + $I_{Top\text{-}K}$
	I_5	F_4 + Top-K	0.460	A + F + S + $I_{Top\text{-}K}$
	I_3	F_4 + CAAF	0.459	A + F + S + I_{CAAF}
	I_7	F_8 + CAAF	0.443	A + F + I_{CAAF}

Table 2: Description of the abbreviation in our method Introduction

Abbreviation	Description	Abbreviation	Description
A	Action recognition	R	Ranking aggregation
F	Face recognition	$I_{Top\text{-}K}$	Top-K feedback
S	Score temporal extension	I_{CAAF}	CAAF feedback

[†]These authors contributed equally to this work.
[*]Corresponding author.

Figure 1: An example for retrieval (programme material copyrighted by BBC)

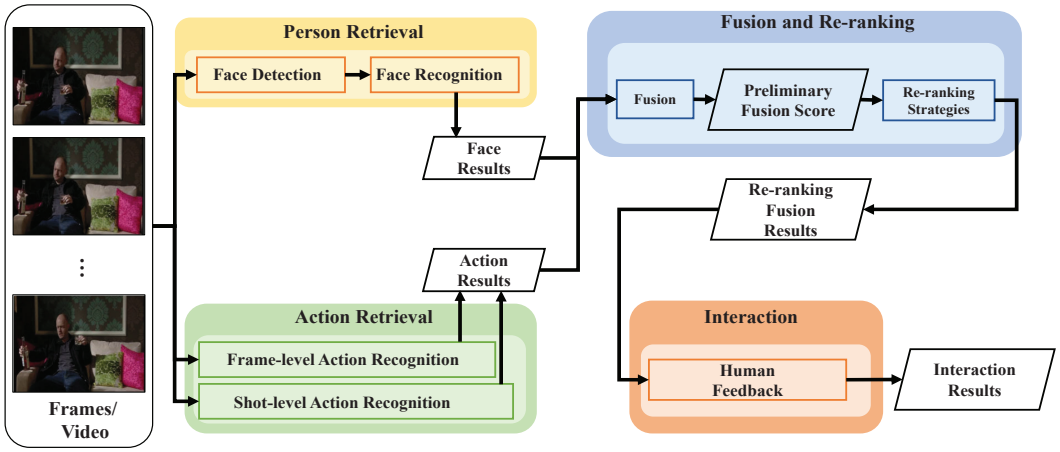

Figure 2: Our framework

1 Introduction

The task of INS [1] in 2021 is retrieving specific person doing specific action as shown in Fig. 1, which can be represented as <person, action>. The dataset has a total of about 464 hours of videos, 471,527 shots, including 191 people, and 25 actions. The 20 topics in 2021 contain 8 actions and 6 characters. Our method is to split the person-action instance search (P-A INS) into two parts: person INS and action INS, and then fuse their separate ranking lists to generate the final results. In addition, this year we also test two interactive methods based on the fusion results.

2 Our Method

As shown in Fig. 2, the framework we proposed for automatic task consists of four parts. The first one is the person retrieval module, which includes face detection using RetinaFace [2] and face recognition using ArcFace [3]. The second part is the action retrieval module, which includes frame-level human object interaction (HOI) detection methods using PPDM [4], QPIC [5] and ASNet [6], and video-level action detection methods using TSM [7] and ACAM [8]. The above two modules are used to generate the person results and the action results respectively, and the third module is the score fusion module for obtaining the ranking list by fusing the above two results. The fourth module is the re-ranking module, which adjusts the ranking result based on the scores and order of the ranking list of the fusion module.

2.1 Person retrieval

The person's identity is achieved by the person retrieval module. To get the efficiency-accuracy trade-off, we resample the input videos at 5 fps. In the experiment, since the clarity of the official face examples is unsatisfactory, we additionally use a self-built face dataset. The official report last year showed that our face dataset performed well and improved the robustness of face recognition, so we continue to use it this year. In the person retrieval module, the first stage is the face detection, in which we choose RetinaFace. Compared with MTCNN [9] used in last year, it runs faster and has higher accuracy. In particular, it improves the detection accuracy when the face rotates at a large angle, which increases the recall rate. Then in the face recognition stage, we choose ArcFace to extract features. This model proposes a new loss function to maximize the classification boundary in the annular space. Compared with the previous Center Loss [10], it is easier to train and has a better classification effect. After that, we compute the similarity matrix by using features from the last step and features from a self-built face dataset to get the retrieval scores.

2.2 Action retrieval

In this module, we use two types of methods. The first one is called frame level method, which can directly detect the interaction between people and objects in the frame, includes PPDM, QPIC and ASNet which are suitable for actions with little or slow change in time. The other type is video level method that can extract spatio-temporal information in the video, including TSM and ACAM, which deals with complex temporal actions well such as opening door and entering.

2.2.1 Frame level

Frame level detection methods include PPDM, QPIC and ASNet, which are used to detect HOI actions in the picture. Compared with general action detection, HOI detection focuses more on detecting the interactive actions of people and obviously objects in the picture, which can better handle some topics with HOI. Among them, PPDM is based on the CNN network structure and performs a single-stage action detection which is proved to be effective in last year. At the same time, we introduced QPIC and ASNet, which are based on Transformer and can make better use of global context information than CNN networks.

2.2.2 Video Level

However, there are still some actions that do not have interaction with objects or cannot be detected through a single frame. To solve these problems, we use TSM and ACAM to extract the spatiotemporal features of the videos. Among them, TSM exchanges channels between the features of neighbouring frames, which can extract video features under the cost of the 2D-CNN model. And ACAM applies I3D [11] on the raw videos to extract features, then use an attention module to capture region of interest (RoI) for action detection, which enhances the effect of the detection with lower time cost.

In the experiment, the score of action is generated by model results directly. For actions which are not in the pre-trained dataset, the model is used to extract the features of the action. Then, the features are compared with the features of the sample action videos, and a similarity matrix is generated to get the retrieval scores.

2.3 Result fusion

After obtaining the respective ranking lists of person and action, the final ranking list is obtained from the result fusion. In this part, since two methods, weight fusion and filter fusion, were verified last year and the second one shows significant advantages in the official report, all submissions this year are based on the filter fusion method. Specifically, we filter the shots containing the target person, and then finetune the action scores on these shots, which can be described as:

$$s_{i,j} = \boldsymbol{F}_{\delta}(Conf_i^{face}) \times Conf_j^{act} \tag{1}$$

$$\boldsymbol{F}_{\delta}(x) = \begin{cases} 1, & x \geq \delta \\ 0, & x < \delta \end{cases} \tag{2}$$

where $s_{i,j}$ means the fusion result of i-th person and j-th action, $Conf_i^{face}$ and $Conf_j^{act}$ is the confidence score of i-th person and j-th action respectively. $\boldsymbol{F}_{\delta}(\cdot)$ is used to calculate whether the face confidence is over threshold and δ is the face threshold. For the shots with the face score less than δ, the fusion score is set to 0, otherwise the action score.

By analyzing the results of each module, we find that there are still many methods to further improve the retrieval accuracy in reasonable time and computational cost. So we plus a re-ranking module additionally.

Take person INS branch as an example, it's easy to find that the face detector performs not well enough when the person has his back or side-to-side towards the camera. Besides, during one shot, the face may be temporarily obscured due to the movement of people or other objects. Similarly, the action also suffers from failed detection when objects which are part of action are temporarily invisible due to person movements or camera scope. So we propose an inter-frame detection extension (IDE) [12], which can fill in the gaps among shots due to detection failures. Firstly it will scan the results of a shot and find the failure slices between two effective detections that has the same person id or action id. Then the confidence score for the face detection box can be complemented as:

$$Conf_i^k = \frac{n}{m+n} \times Conf_i^{k-m} + \frac{m}{m+n} \times Conf_i^{k+n} \tag{3}$$

where $Conf_i^k$ is the confidence score of i-th person in the k-th keyframe which failed to detect. $Conf_i^{k-m}$ and $Conf_i^{k+n}$ are the scores of its neighbours, the $(k-m)$-th and $(k+n)$-th keyframes. We use linear interpolation to calculate the confidence scores and the positions of face detection boxes.

However, directly fusing the results of person and action module like Eq. (1) is suboptimal, because of $identity\ inconsistency\ problem$ (IIP), which means that there may be more than one person in one shot, and the person identity of detected face and action from two branches may not accordant. Hence, we propose identity consistency verification (ICV) [12] to solve the problem. Since both action and person module can provide the box of person, we can calculate the intersection over union (IoU) of the action box and face box. The assumption is that the action box and face box will have a higher possibility of belonging to the same owner if they have higher IoU, and Eq. (1) can be rewritten as:

$$s_{i,j} = c_{i,j} \times \boldsymbol{F}_{\delta}(Conf_i^{face}) \times Conf_j^{act} \tag{4}$$

$$c_{i,j} = \frac{\boldsymbol{Area}(Box_i^{face} \bigcap Box_j^{act})}{\boldsymbol{Area}(Box_i^{face})} \tag{5}$$

where $s_{i,j}$ is the score of i-th person and j-th action in the keyframe. $c_{i,j}$ is the score of ICV result which is added as weight of origin result. Box_i^{face} and Box_j^{act} is the box of i-th person and j-th action respectively. We use the percentage of the overlapping area in the face area as $c_{i,j}$.

2.4 Re-ranking

Additionally, we found that some actions have temporal continuity and can last more than one shot, so we propose score temporal expansion (STE) [12], which adjusts the fusion score of the special shots by fusing the scores of neighbour shots. In the experiment, we set the diffusion direction from higher confidence shots to lower ones, and calculate the score according to the difference between the score of two shots and the distance between two shots:

$$s_{i,j}^{k\_ste} = s_{i,j}^{k\_ori} + \theta \sum_{-p<m<p} \boldsymbol{F}_{dis}(m) \times \mathbf{max}(s_{i,j}^{(k+m)\_ori} - s_{i,j}^{k\_ori}, 0) \tag{6}$$

$$\boldsymbol{F}_{dis}(m) = e^{-\frac{m^2}{\sigma}} \tag{7}$$

where $s_{i,j}^{k\_ste}$ is the revised score of i-th person and j-th action in the k-th shot after STE stage, $s_{i,j}^{k\_ori}$ is original score, and $F_{dis}(\cdot)$ is the distance weight decaying with the length between two shots. Using maximum function to limit direction and hyperparameter θ and σ to change size.

The Ranking Aggregation (RA) strategy can obtain better results than any of the original ranking lists by fusing the results. We apply the method in [13], which is based on the Half-Quadratic (HQ) theory and optimizes the results by minimizing the distance between the result sorting and all input sorting. Different from the traditional sorting fusion algorithm, this algorithm uses the HQ function instead of Euclidean distance to measure the distance, so that the final result is less affected by abnormal sorting. At the same time, based on the HQ theory, the algorithm transforms the distance minimization problem into an iterative problem. Through continuous iteration until the weight vector converges, the weight of each order is calculated and merged to obtain the final order. The objective function of the algorithm is as follows:

$$\min_{R^*,\alpha} J(R^*, \alpha) = \sum_{m=1}^{M} \alpha_m \left\| R^m - R^* \right\|_2^2 + \psi(\alpha_m) \tag{8}$$

Among them, M is the number of sorted lists, R^m is the sort of the mth sorted list, R^* is the total sort of the aggregated M sorted lists, $\alpha \in R^M$ is the HQ auxiliary variable, and $\psi(\cdot)$ is the conjugate convex function of the HQ function $g(\cdot)$.

2.5 Interaction

In the interaction module, we used two different interaction methods called Top-K Feedback and confidence-aware activate feedback (CAAF) [14], both of which are proved to be effective.

2.5.1 Top-K Feedback

For the first method, we used a simple rearrangement method that is directly labeling the top-k shots in the ranking list firstly, then promoting the marked positive samples to the top of the list and putting the negative examples to the end. Since this method does not introduce the wrong samples to the top of the list, the experimental results show that the method performs well. Besides, we have many labeling strategy for different conditions.

- Only label positive examples: Successively check top-K examples and only label positive ones, then put them to the top of ranking list.
- Only label negative examples: Successively check top-K examples and only label negative ones, then put them to the bottom of ranking list.
- Label positive and negative examples: Successively check top-K examples and label both positive and negative ones, the positive examples of that are pull up to the top while the negative examples of that are push down to the bottom.

we can only label positive or negative examples for speeding up or label both of them for fine-grained labels.

2.5.2 CAAF

Considering that the features of similar actions are closer in feature space, while different actions are farther away, we apply CAAF. This method can dynamically recommend elements according to the current ranking list and the interaction condition of the elements. CAAF chooses the average of the top $A(A > 10)$ results as probe. Each elements of probe and gallery has a confidence score represented by v and ranking score represented by f. CAAF chooses high quality samples according to v, and changes the ranking list according to f. Human feedback can change v, and CAAF can correspondingly refresh f for all elements. The optimization goal is shown below:

$$\min_{\boldsymbol{f},\boldsymbol{v}} \mathcal{E}(\boldsymbol{f},\boldsymbol{v}) = \mathcal{L}(\boldsymbol{f},\boldsymbol{v}) + \mathcal{R}(\boldsymbol{v}) \tag{9}$$

Where i and j means the i-th and j-th element in the set X which includes both probe and gallery elements. $\mathcal{L}(\boldsymbol{f},\boldsymbol{v})$ is calculate by $\mathcal{L}(\boldsymbol{f},\boldsymbol{v}) = \frac{1}{m^2}\sum_{i,j}(v_i + v_j)(l_{ij} - \beta)$, and $\mathcal{R}(\boldsymbol{v})$ is a penalty for limit v. l_{ij} and β) are pairwise loss and loss threshold.

3 Analysis

Table. 1 shows the mAP and method used in the all of our submission, so we analyze the effectiveness of each technique by comparing the difference between them. By only fusing the score of action and person module and using PPDM, QPIC, TSM and ACAM for action module, F_8[‡] is our automatic baseline that achieves 0.395 mAP. F_6 and F_4 are equipped with Score Extension strategy, both of which get the mAP boost of 0.023 mAP . The difference of them is the choice of topics, for we only use Score Extension on the action which may last for more than one shot.

[‡]F_8 is the abbreviation of the submission ID F_M_A_B_WHU_NERCMS.21_8, for the interactive submissions, I_7 is the abbreviation of I_M_A_B_WHU_NERCMS.21_7, we will use the similar form below.

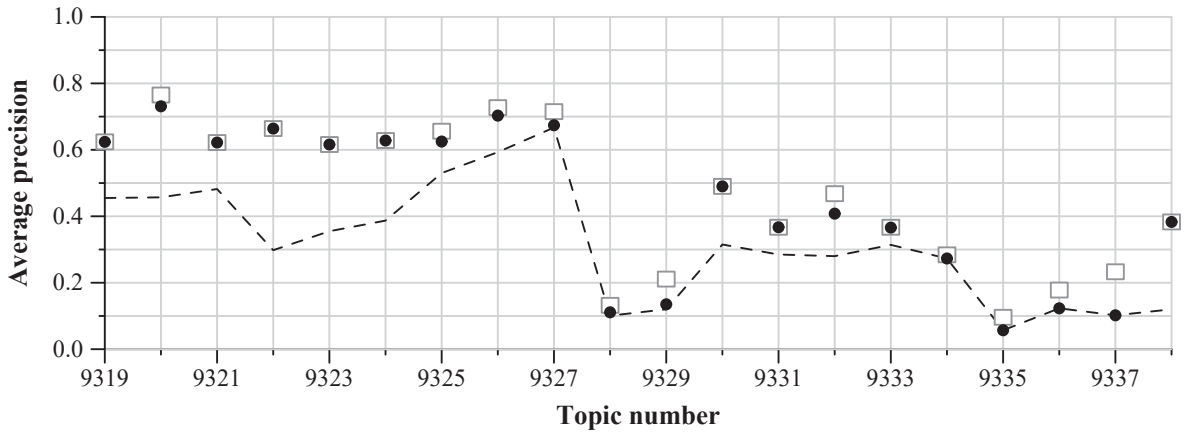

Figure 3: Comparison of automatic task submission

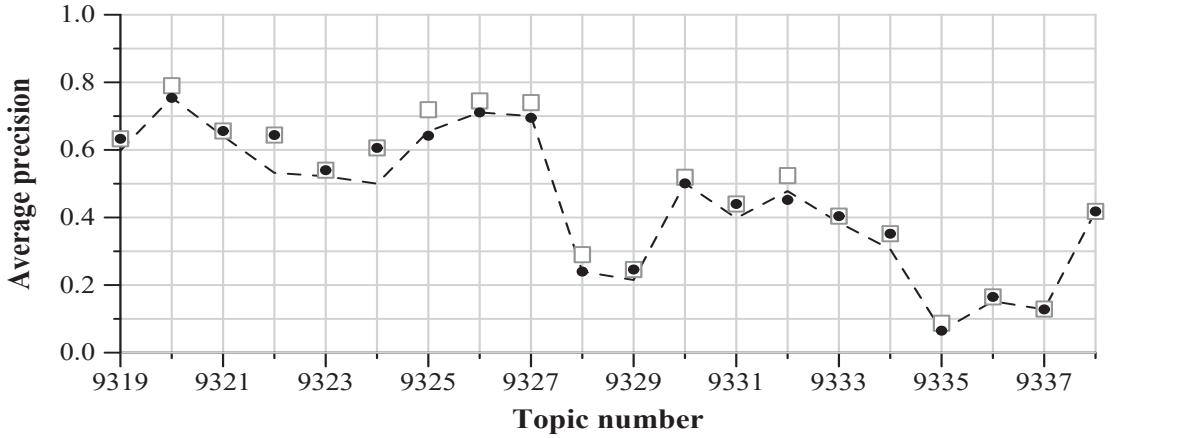

Figure 4: Comparison of interactive task submission

However, the constancy of some action are alterable according to the context. And the result reveals that Score Extension performs not well enough on these actions. F_2 is added with Ranking Aggregation strategy, which contributes 0.017 mAP. We use PPDM, QPIC and ASNet to generate result respectively, and after that, use the ensemble method in [13] to fuse the above three score to generate a better ranking list. So it is the best automatic run among our submissions.

For interactive task, I_7, the baseline of our interactive submission, uses CAAF strategy on I_8. With Query Expansion strategy, CAAF can find the valuable example for human interaction so that it can get 0.048 mAP increase. I_5 and I_3 apply Top-K Feedback and CAAF on I_4 and got 0.042 and 0.041 mAP increase respectively. Since Top-K Feedback can provide more interactive examples and CAAF can provide more valuable examples, these submissions achieve the comparable result. I_1 is generated by Top-K Feedback strategy from I_2, due to higher automatic result, it is the best interactive run among our submissions and up 0.030 mAP.

Fig. 3 and Fig. 4 shows the automatic and interactive result of I_2 and I_1 in all topic this year. The 'dot', 'box' and '-' are our score, the best score and the median score of each submissions from all entrants. It is obvious that our

best submission achieves top-1 at quite a lot of topics (9 for automatic task and 12 for interactive task). For many topics with obvious HOI, our method exhibits superior accuracy, due to the advanced frame level detector. However, in some topics, such as 9328 (<'Max', 'carrying bag'>) and 9329 (<'Peggy', 'carrying bag'>), our method performs not well on them. By reviewing the ranking lists, we find that our model prefer to mistake 'bag' for 'belt', which is similar with necklace and other jewelry and hence generates wrong results. For 9335 (<'Bradley', 'open door enter'>) and 9336 (<'Pat', 'open door enter'>), the action changes a lot over time and is confusing compared with 'open door leave', 'open door stand' and so on. So our method does not perform well on these actions.

4 Conclusion

Through the INS task in TRECVID 2021, we conduct extensive experiments for our framework. By using advanced models and re-ranking strategies, our submission achieves the 1^{st} place for automatic and interactive searches.

Acknowledgement

This work is supported by National Nature Science Foundation of China (No. U1903214, 61876135, 62071338, 61862015), National Nature Science Foundation of Hubei Province (2019CFB472) and Hubei Province Technological Innovation Major Project (2018AAA062). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] George Awad, Wessel Kraaij, Paul Over, and Shin'ichi Satoh. Instance search retrospective with focus on trecvid. *International Journal of Multimedia Information Retrieval*, 6(1):1–29, 2017.

[2] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019.

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[4] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490, 2020.

[5] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021.

[6] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9004–9013, 2021.

[7] Ji Lin, Chuang Gan, Kuan Wang, and Song Han. Tsm: Temporal shift module for efficient and scalable video understanding on edge devices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[8] Oytun Ulutan, Swati Rallapalli, Mudhakar Srivatsa, Carlos Torres, and BS Manjunath. Actor conditioned attention maps for video action detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 527–536, 2020.

[9] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.

[10] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, 2016.

[11] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[12] Jingyao Yang, Chao Liang, Yanrui Niu, Baojin Huang, and Zhongyuan Wang. A spatio-temporal identity verification method for person-action instance search in movies. *arXiv preprint arXiv:2111.00228*, 2021.

[13] Majid Mohammadi and Jafar Rezaei. Ensemble ranking: Aggregation of rankings produced by different multi-criteria decision-making methods. *Omega*, 96:102254, 2020.

[14] Yue Zhang, Chao Liang, and Longxiang Jiang. Confidence-aware active feedback for efficient instance search. *arXiv preprint arXiv:2110.12255*, 2021.