# Human-in-the-Loop Vehicle ReID

**Zepeng Li**[1], **Dongxiang Zhang**[1*], **Yanyan Shen**[2], **Gang Chen**[1]

[1] Key Lab of Intelligent Computing Based Big Data of Zhejiang Province, Zhejiang University
[2] Department of Computer Science and Engineering, Shanghai Jiao Tong University
{lizepeng,zhangdongxiang,cg}@zju.edu.cn, shenyy@sjtu.edu.cn

## Abstract

Vehicle ReID has been an active topic in computer vision, with a substantial number of deep neural models proposed as end-to-end solutions. In this paper, we solve the problem from a new perspective and present an interesting variant called human-in-the-loop vehicle ReID to leverage interactive (and possibly wrong) human feedback signal for performance enhancement. Such human-machine cooperation mode is orthogonal to existing ReID models. To avoid incremental training overhead, we propose an Interaction ReID Network (IRIN) that can directly accept the feedback signal as an input and adjust the embedding of query image in an online fashion. IRIN is offline trained by simulating the human interaction process, with multiple optimization strategies to fully exploit the feedback signal. Experimental results show that even by interacting with flawed feedback generated by non-experts, IRIN still outperforms state-of-the-art ReID models by a considerable margin. If the feedback contains no false positive, IRIN boosts the mAP in Veri776 from $81.6\%$ to $95.2\%$ with only $5$ rounds of interaction per query image.

## Introduction

Given a query image and an image gallery harnessed across multiple surveillance cameras, vehicle ReID retrieves images that refer to the same real-world vehicle. The problem is challenging due to the presence of different viewpoints (Lou et al. 2019), low-image resolution (Zhao et al. 2021), illumination changes (Liu et al. 2016a) and partial occlusions (Rao et al. 2021; Zhang et al. 2022). To overcome these challenges, state-of-the-art methods typically resort to devising advanced neural networks (Rao et al. 2021; Zhao et al. 2021) or effective loss functions (Yan et al. 2020; Quispe et al. 2021) to extract discriminative visual features and establish remarkable performance in the benchmark datasets. Nonetheless, new breakthrough via model improvement has become more and more challenging.

In this paper, we jump out of the box and propose a new paradigm called human-in-the-loop vehicle ReID. It is orthogonal to existing ReID models and works in a human-machine cooperative mode to leverage iterative feedback for further performance improvement. Note that the quality of feedback could be unreliable, because identifying two matching vehicles through their appearance is also challenging for human beings[1]. Figure 1 depicts a toy example to deliver the overall idea. In the initial step, a list of images is returned and ranked by the similarity to the query image, using any existing ReID models. Afterwards, users are allowed to employ an operation to provide feedback signal on the results. To avoid incurring cumbersome efforts for human intervention, we define the operation as picking a positive match from a small set of uncertain candidates. Our goal is to develop a mechanism to effectively take advantage of the feedback signal and update the order of images in the rank list. In the next-round interaction, a new subset of uncertain images will be selected for human verification. This process is repeated until the final results are satisfactory or a maximum number of iterations is reached.

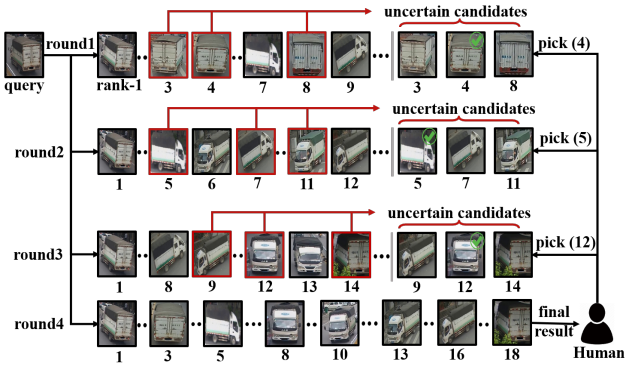



Figure 1: An example of human-in-the-loop vehicle ReID.

A straightforward approach to leverage human feedback is to treat the human-picked positive sample as a new observation and apply incremental learning to update the ReID results. For instance, we can adopt the online incremental learning framework proposed in (He et al. 2020) to maintain an exemplar set for each vehicle class. In each iteration, this set is updated to incorporate new training samples derived from human feedback. The network retains previous knowledge as part of the final classifier and is re-trained us-

[*] Corresponding author



[1] The annotation of vehicle ReID benchmark datasets often needs to rely on additional clues such as vehicle plate or travel routes.

ing the exemplar set. (Wu and Gong 2021) designs a more comprehensive learning objective that incorporates the coherence of classification, distribution and representation in a unified framework. The underlying motivation is to support life-long ReID without forgetting. However, these incremental learning methods suffer from two major drawbacks when applied in the online and interactive scenario of Figure 1. First, the human feedback in each iteration is lightweight as it only contains a "positive" sample selected from an uncertain candidate set. Its effect is limited even with the assistance of data augmentation. Second, it incurs additional online training cost, which is not friendly for real-time human-machine interaction.

To resolve these two issues, we propose a novel mechanism to better exploit the human feedback signal without incurring additional training cost. Our idea is to devise a network that accepts human feedback as part of the input and dynamically adjusts the embedding of query image to reduce its distance to the positive samples. To achieve the goal, we propose an Interaction ReID Network (IRIN), which augments the existing ReID models with a Transformer (Wu et al. 2021) encoder with gating mechanism. In the offline training of IRIN, we simulate the human interaction process to generate the feedback signal, by constructing a selective uncertain candidate set for each vehicle in the training set and picking a positive sample using the groundtruth labels. In addition, we adopt supervised contrastive learning (Khosla et al. 2020) to pull the query image and its positive samples closer in the embedding space. Finally, IRIN is jointly trained with the backbone ReID model to minimize the identification loss and contrastive loss.

We invite 20 postgraduates with 10 female students and 10 male students for performance evaluation. It is possible that these students pick false positives from the uncertain candidate sets and generate misleading feedback signals. Experimental results show that the mAP of IRIN increases steadily with more iterations, implying that it can effectively leverage the feedback signal. Even when IRIN is provided with flawed feedback, the new human-machine cooperation mode can still surpass pure machine models, student-based annotations and existing online learning frameworks. In the ideal case with perfect feedback, with only 5 rounds of interaction for each query image, IRIN boosts the mAP in Veri776 (Liu et al. 2016b) from $81.6\%$ to $95.2\%$. In the end of the experiments, we also apply IRIN in the task of person ReID to validate its generality.

## Related Work

**Vehicle ReID:** The mainstream strategy of vehicle ReID is to learn robust and discriminative vehicle representation via devising advanced neural network or effective loss functions. In the former category, (Zhao et al. 2021) proposes a heterogeneous relational complement network that combines region-specific features and cross-level features as a supplement to the original high-level output. (Khorramshahi et al. 2019) employs a dual adaptive attention mechanism to focus on the most informative key-points of vehicle image. (He et al. 2021) proposes a Transformer-based ReID model, with a jigsaw patch module and side information embeddings to enhance the robust feature learning. The model achieves state-of-the-art performance in vehicle ReID benchmark datasets. As to loss function improvement, Circle loss (Sun et al. 2020) is developed to achieve a more flexible and targeted pair similarity optimization. To stabilize the triplet loss, (Ghosh, Shanmugalingam, and Lin 2021) proposes a grid-based motion statistical feature matcher for relation-preserving triplet mining. (Quispe et al. 2021) uses attribute-related cross entropy loss and triplet loss to distill crucial attribute information. Readers can refer to (Zakria et al. 2021) for a comprehensive survey.

**Human-in-the-Loop Visual Tasks:** We review human-in-the-loop visual tasks according to the machine learning pipeline, including the stages of data annotation, model training and online inference. 1) **Data annotation**: To improve data quality, (Berti-Équille 2019; Muthuraman et al. 2021) utilize model sensitivity to identify potentially incorrect labels for human verification. In (Liu et al. 2019), reinforcement learning is adopted in the task of person ReID to iteratively prioritize a set of data samples for human annotation. 2) **Model training**. This step is focused on how to iteratively leverage human feedback to improve model performance. As mentioned, (He et al. 2020) and (Wu and Gong 2021) are two representative incremental learning strategies that work in the online scenario. 3) **Online inference**: In this stage, human can assist models to accomplish a task together and achieve better performance. For instances, (Brenner, Priyadarshi, and Itti 2016) leverage the human feedback of online viewpoint adjustment to improve object detection confidence without additional training cost. (Stonebraker et al. 2020) presents a search-and-mark framework to facilitate surveillance tasks.

This work belongs to the stage of online inference, i.e., human users work with ReID models in an online fashion, without the need to re-train the underlying model. The research challenges come from the requirements for real-time interaction and robustness to the flawed feedback.

## Proposed Model

### Framework Overview

In this paper, we study iterative vehicle ReID with the assistance of human feedback for continuous performance improvement. Given a query image $I_q$, we define the human operator as picking the most promising image $I_p$ from a set of uncertain candidates $U$ and represent the feedback as a positive matching pair $(I_q, I_p)$. Other types of human operator can also be explored and we consider this direction as our future work. With the feedback signal, our idea is to devise a network that can accept the human feedback as part of input and dynamically adjust the embedding of query image to reduce its distance to the positive samples. As shown in Figure 2, we propose an Interaction ReID Network (IRIN) that augments a backbone ReID model with a Transformer encoder with gating mechanism. In the offline training stage, we simulate the human interaction process to generate the feedback signal, by constructing a selective uncertain candidate set for each vehicle in the training set and picking a positive sample using the groundtruth labels. In addition,

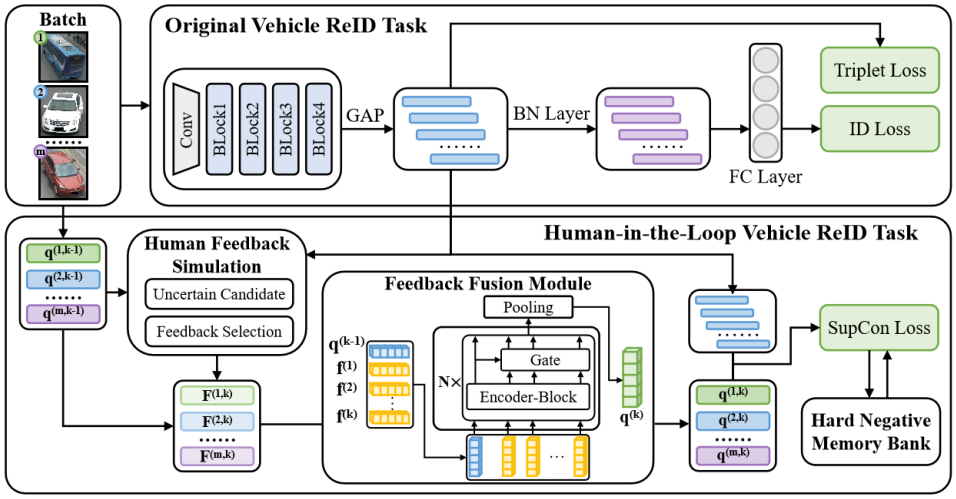



Figure 2: The network structure and training pipeline of IRIN.

we treat the query image as an anchor and adopt supervised contrastive learning (Khosla et al. 2020) to pull together the anchor and its positive samples in the embedding space. Finally, IRIN is jointly trained to minimize the identification loss and contrastive loss.

## Backbone ReID Network

We adopt ResNeXt101 (Xie et al. 2017) appended by an Instance Batch Normalization (IBN) (Pan et al. 2018) layer as our backbone network, mainly due to its simplicity and popularity — it is easy to implement and has been widely applied in previous ReID models (Zhao et al. 2021; Ghosh, Shanmugalingam, and Lin 2021). Our IRIN is built on top of the backbone network and augment it with a module to fuse the human feedback and dynamically adjust the feature of query image.

## Human Feedback Simulation

Since IRIN treats the human feedback signal as part of the input, we need to simulate the procedure of human interaction in the offline training state. In our simulation process, we randomly pick an image from each vehicle in the training set to constitute a query set $Q$. For each $I_q \in Q$, we determine a group of uncertain samples $U$ that require human assistance. More specifically, we adopt the popular sampling strategy in active learning and select uncertain samples based on the classification entropy of the model output.

$$x_U = \arg\max_x -\sum_y P_\theta(y \mid x) \log P_\theta(y \mid x) \qquad (1)$$

where $x$ refers to a vehicle image, $y$ is its class id, and $P$ is the vehicle classification probability. $\theta$ is our model parameters.

Since our human operator is defined as picking a matching image from a set of uncertain candidates, we need to select an instance $I_p$ from $U$ to generate human feedback and the instances in $U$ that belong to the same vehicle as the query image are beneficial to improve the performance of the model against the query. Intuitively, $I_p$ with the maximum distance to $I_q$ is preferred because it is the hardest sample that cannot be well resolved by the current model. However, we observe that users are inclined to select the sample that they feel the most promising, i.e., with the highest probability of belonging to the same vehicle as the query. This observation motivates us to simulate human behavior by selecting $I_p$ from set $U$, which is associated with the *minimum* distance to $I_q$.

Our offline training process also simulates another feature of human-in-the-loop ReID, in which the iterative feedback is applied on the same query image for a number of consecutive iterations. In our batch setting, let $B$ be the batch size in model training. In each batch, we select $m$ vehicles and each vehicle is assigned with $\frac{B}{m}$ training samples. In the subsequent iterations, we will train the model with only the samples from these $m$ vehicles until all of their samples have been accessed. Afterwards, we proceed to the next group of $m$ vehicles.

## Feedback Fusion Module

We propose a Transformer-based encoder with gating mechanism to fuse the query feature with the feedback signals. In the $k$-th iteration, the feedback signal $\mathbf{f}^{(k)}$ (i.e., the visual feature of the picked image) is concatenated with previous signals to form a matrix $\mathbf{F}^{(k)} = [\mathbf{f}^{(1)}, \mathbf{f}^{(2)}, \ldots, \mathbf{f}^{(k)}]$. Let $\mathbf{q}^{(k-1)}$ denote the fused query feature in the $(k-1)$-th iteration. Our goal is to fuse $\mathbf{q}^{(k-1)}$ and $\mathbf{F}^{(k)}$ to derive a new query-specific feature $\mathbf{q}^{(k)}$. In our implementation, $\mathbf{q}^{(k-1)}$ is first concatenated with $\mathbf{F}^{(k)}$ and used as the input to the Transformer-based encoder. We use $[\mathbf{q}_1^{(k-1)}, \hat{\mathbf{f}}_1^{(1)}, \ldots, \hat{\mathbf{f}}_1^{(k)}]$ to denote the output by the first encoder layer $\mathrm{E}_1$. Since the feedback signals could be false positive, we also devise a gating mechanism to determine the contribution of the encoded feature $\hat{\mathbf{f}}_1^{(i)}$. In particular, $\mathbf{q}_1^{(k-1)}$ and $\hat{\mathbf{f}}_1^{(i)}$ $(i = 1, \ldots, k)$ are concatenated to calculate the gate weight

through a multi-layer perceptron (MLP) and activate function (Sigmoid). If the feedback refers to a positive match, we expect the weight output by Sigmoid to be close to 1. Otherwise, a small gate weight is preferred. As a trade-off between efficiency and model performance, we use 2-layer encoder for the feature fusion between $\mathbf{q}^{(k-1)}$ and $\mathbf{F}^{(k)}$. Finally, $\mathbf{q}^{(k)}$ is obtained by pooling the output of the two-layer encoder.

$$[\mathbf{q}_1^{(k-1)}, \hat{\mathbf{f}}_1^{(1)}, \ldots, \hat{\mathbf{f}}_1^{(k)}] = \mathrm{E}_1([\mathbf{q}^{(k-1)}, \mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(k)}]) \quad (2)$$

$$\mathbf{f}_1^{(i)} = \hat{\mathbf{f}}_1^{(i)} \odot \mathrm{Sigmoid}(\mathrm{MLP}_1(\mathbf{q}_1^{(k-1)} \oplus \hat{\mathbf{f}}_1^{(i)})) \quad (3)$$

$$[\mathbf{q}_2^{(k-1)}, \hat{\mathbf{f}}_2^{(1)}, \ldots, \hat{\mathbf{f}}_2^{(k)}] = \mathrm{E}_2([\mathbf{q}_1^{(k-1)}, \mathbf{f}_1^{(1)}, \ldots, \mathbf{f}_1^{(k)}]) \quad (4)$$

$$\mathbf{f}_2^{(i)} = \hat{\mathbf{f}}_2^{(i)} \odot \mathrm{Sigmoid}(\mathrm{MLP}_2(\mathbf{q}_2^{(k-1)} \oplus \hat{\mathbf{f}}_2^{(i)})) \quad (5)$$

$$\mathbf{q}^{(k)} = \mathrm{Pooling}([\mathbf{q}_2^{(k-1)}, \mathbf{f}_2^{(1)}, \ldots, \mathbf{f}_2^{(k)}]) \quad (6)$$

## Supervised Contrastive Learning

Contrastive learning is a self-supervised approach to learn an embedding space in which similar sample pairs are pulled together while dissimilar ones stay far apart. Supervised contrastive learning extends the self-supervised mode into fully-supervised setting to effectively leverage label information. It chooses positive pairs from the same class and negative samples from different classes. The learning objective is to pull together samples belonging to the same class and push apart those from different classes. To utilize supervised contrastive learning, we maintain the query feature as anchor, which will be iteratively updated. Images referring to the same vehicle with the query vehicle will be regarded as positive samples while others as negative samples. Considering the quantitative limitation of batch size on negative samples, we propose a memory bank called Hard Negative Memory Bank (HNMB) for negative sample expansion with at low cost inspired by (Wu et al. 2018). More specifically, we maintain a fixed-size queue to store negative samples for each query vehicle, which is denoted by $N^+(i)$. Hard negatives in the batch samples are continuously stored in $N^+(i)$ and follow the principle of first-in-first-out for eviction, i.e., the earliest samples will be popped as long as the size of $N^+(i)$ exceeds the memory size.

## Joint Training

As shown in Figure 2, the feedback fusion module is jointly trained with the backbone ReID network, with two optimization objectives. First, for the backbone network, we can simply apply the loss function commonly used in previous ReID models to learn discriminative feature representation. In our implementation, we follow recent works (Ghosh, Shanmugalingam, and Lin 2021; Quispe et al. 2021) to adopt the combination of ID Loss and Metric Loss. As shown in the following equations, we choose Cross Entropy Loss as the ID Loss and soft-margin Triplet Loss as the Metric Loss:

$$\mathcal{L}_{ID} = \sum_{i=1}^{N} -\hat{\mathbf{p}}_i \log(\mathbf{p}_i), \quad \begin{cases} \hat{\mathbf{p}}_i = 0, i \neq y \\ \hat{\mathbf{p}}_i = 1, i = y \end{cases} \quad (7)$$

$$\mathcal{L}_{Metric} = \log[1 + \exp(\|\mathbf{v}_a - \mathbf{v}_p\| - \|\mathbf{v}_a - \mathbf{v}_n\|)] \quad (8)$$

where $y$ is the groundtruth ID label, $\mathbf{p}_i$ is the ID prediction logits of class $i$, $\mathbf{v}_a$ is an anchor feature, $\mathbf{v}_p$ is a positive feature and $\mathbf{v}_n$ is a negative feature.

Second, for the component of feedback signal fusion, we apply Supervised Contrastive Loss (SCL) $\mathcal{L}_{SCL}$ to adjust the query embedding using feedback and make it as close as possible to those of the matching images. We apply SCL to this task because it encourages the model to pay more attention to the hard samples (containing both positives and negatives) so as to generate a more discriminative query embedding by leveraging feedback signals. $\mathcal{L}_{SCL}$ can be formally expressed as:

$$\mathcal{L}_{SCL} = \sum_{i \in I} \frac{-1}{N} \sum_{p \in P(i)} \log \frac{\exp\left(\mathbf{q}^{(k)} \cdot \mathbf{v}_p^\top / \tau\right)}{\sum_{a \in A(i)} \exp\left(\mathbf{q}^{(k)} \cdot \mathbf{v}_a^\top / \tau\right)} \quad (9)$$

where we set the iteratively updated vehicle feature $v_{q,k}$ as anchor in SCL. $A(i) = P(i) + N(i) + N^+(i)$ is a set of vehicle image indices in the batch for vehicle $i$, in which $P(i)$ is the set of indices of all positives, $N(i)$ is all negatives indices set, and $N^+(i)$ is HNMB for vehicle $i$ after $k-1$ iterations of simulated human feedback.

With the two types of loss functions, the final objective of joint learning is to minimize

$$\mathcal{L} = \lambda_{id}\mathcal{L}_{ID} + \lambda_m\mathcal{L}_{Metric} + \lambda_h\mathcal{L}_{SCL} \quad (10)$$

where $\lambda_{id}$, $\lambda_m$ and $\lambda_h$ are weight parameters.

## Online Inference

With the trained IRIN model, we explain the online inference procedure with iterative feedback signals (as shown in Algorithm 1). Given a query image, we extract its features using the backbone network (which is ResNeXt101 appended by an IBN in our implementation). The images in the gallery are sorted according to their similarity to the query image and top-$n$ candidates are returned. Among the returned images, a set of uncertain images $U$ are identified according to the classification entropy. The user picks an image $I_p$ from $U$ that is considered to be positive with the highest confidence. With the feedback signal, the query feature is adjusted by the IRIN network. In the next iteration, the top-$n$ candidates and uncertain image set $U$ are updated. The user picks another "positive" image from $U$ to update query feature. The procedure repeats until the maximum number of iterations is reached or the user is satisfied with the top-$n$ results.

# Experiment

## Experimental Setup

**Datasets**. We use two popular vehicle ReID benchmarks. **Veri-776** (Liu et al. 2016b) contains $49,357$ images of $776$ different vehicles, captured by $20$ cameras in multiple viewpoints. **VehicleID** (Liu et al. 2016a) is a larger-scale dataset, with $221,567$ images and $26,328$ vehicles. In evaluation, it provides three test datasets in different scales (*small*, *medium*, and *large*).

**Implementation Details**. Following previous ReID models, the input images are resized to $240 \times 240$ and augmented

**Algorithm 1:** Vehicle Re-ID with iterative feedback

$\mathbf{q}^{(0)} \leftarrow$ Extract visual feature of the query image;
$CAND_n \leftarrow$ top-$n$ similar images to $\mathbf{q}^{(0)}$;
$U \leftarrow$ images in $CAND_n$ with top-$|U|$ entropy;
**for** $k \leftarrow 1$ **to** $I_{max}$ **do**
  **if** $CAND_n$ *is satisfactory* **then**
    **break**
  $I_p \leftarrow$ the image picked by the user from $U$;
  $\mathbf{q}^{(k)} \leftarrow IRIN(\mathbf{q}^{(k-1)}, I_p)$;
  Update $CAND_n$ and $U$ according to $\mathbf{q}^{(k)}$;
**return** the sorted images w.r.t. the similarity to $\mathbf{q}^{(k)}$

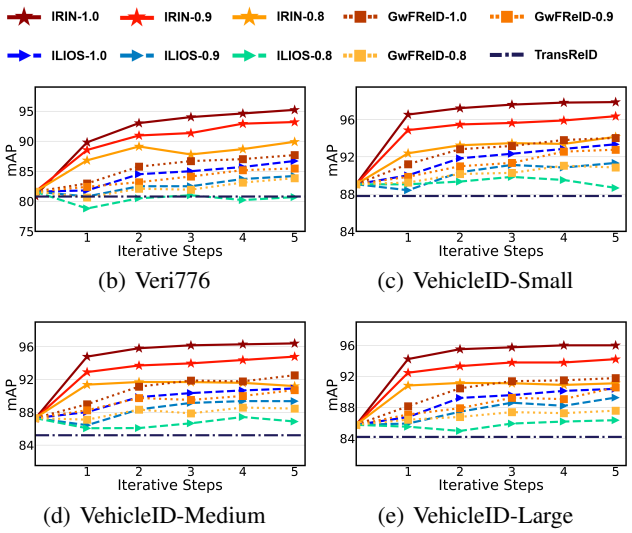
by random flipping, random padding and random erasing. The feature dimension is set to $2,048$. The model is trained with $120$ epochs with a batch size of $128$. SGD optimizer is employed with a momentum of $0.9$ and the weight decay of $5e-4$. Each batch contains $8$ images per vehicle. The initial learning rate is set to $0.01$ and linearly decayed to $0.0001$. Our backbone network is created by appending an Instance Batch Normalization (IBN) (Pan et al. 2018) to a ResNeXt101 (Xie et al. 2017) model and we select $10$ uncertain samples for human interaction simulation. The size of hard negative memory bank is set to $512$. In the loss function of joint training, the weight parameters $\lambda_{id}$, $\lambda_m$ and $\lambda_h$ are both set to $1.0$. In the online inference stage, top-$50$ candidates are returned in each iteration and the size of uncertain images $U$ is set to $10$. The model is implemented with PyTorch and trained on Tesla-V100 GPU.

**Performance Metrics**. The performance of our human-in-the-loop vehicle ReID can still be evaluated by conventional ReID metrics. We select mean average precision (mAP), Cumulative Matching Characteristics at top-1 (rank-1) and Cumulative Matching Characteristics at top-5 (rank-5) for performance evaluation.

**Comparison Methods**. We consider TransReID (He et al. 2021) as the state-of-the-art vehicle ReID model. In our performance evaluation, it is treated as a machine baseline without human intervention. As to incremental learning in the online scenario, we select ILOS (He et al. 2020) and GwFReID (Wu and Gong 2021) as two representative approaches. They maintain an exemplar set for each vehicle which incorporates the candidates picked by the user with high confidence. Online training is conducted on the exemplar set, with different learning objectives.

## Sensitivity to Human Feedback Quality

Since it is possible that users provide flawed feedback with false positives, we simulate the human interaction in the evaluation stage and control the probability that a correct positive pair is picked as the feedback. As shown in Figure 3, we vary the probability of positive feedback from $1.0$ to $0.8$ and compare IRIN with online incremental learning methods (ILOS and GwFReID). The mAP of TransReID is also plotted as the machine baseline without human intervention. It is interesting to find that the backbone model of IRIN can achieve higher mAP than TransReID, probably due to the joint training framework.




(b) Veri776 (c) VehicleID-Small

(d) VehicleID-Medium (e) VehicleID-Large

Figure 3: Sensitivity to the quality of human feedback.

When the feedback signal is relatively reliable, we can see that the mAPs of IRIN, ILOS and GwFReID increase with more rounds of human interaction, indicating that both IRIN and online incremental learning can benefit from iterative human feedback — their mAPs also increase with the quality of feedback. In the oracle scenario with perfect feedback, with only $5$ iterations for each query image, IRIN can boost the accuracy to $95.2\%$ from $81.6\%$ in Veri776, which is remarkably higher than $80.8\%$ achieved by TransReID. IRIN can significantly better leverage the feedback signal than the incremental learning frameworks of ILOS and GwFReID. With the same quality of the feedback, the mAP of IRIN surpasses ILOS and GwFReID by a wide margin. Between the two competitors, GwFReID outperforms ILOS because it is associated with a more comprehensive learning objective.

## Sensitivity to Candidate Set Size

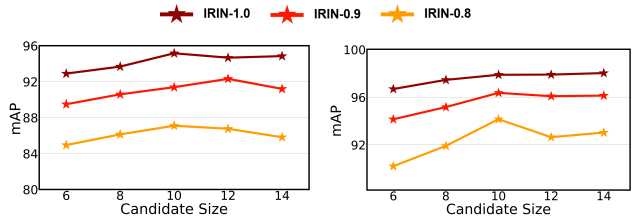
We also investigate how the size of uncertain candidate images impacts human-in-the-loop vehicle ReID. Figure 4 shows that the mAP results for IRIN-1.0, IRIN-0.9 and IRIN-0.8 under different sizes of candidate set. Intuitively, the IRIN-1.0 that always receives a true positive feedback can benefit from the expansion of candidate set, whereas IRIN-0.9 and IRIN-0.8 with noisy feedback cannot increase steadily as the feedback candidate set increases. Since human feedback is not perfect, we set the default size of candidate images to $10$ in our experiments with real human interaction.

## Efficiency Study

In this experiment, we evaluate the time cost spent on human feedback signal processing. IRIN can directly accept the feedback as part of the input and adjust the feature embedding for the query image. Thus, its running time includes

| Method | Veri776 | | VehicleID-Small | | VehicleID-Medium | | VehicleID-Large | |
|---|---|---|---|---|---|---|---|---|
| | mAP | time | mAP | time | mAP | time | mAP | time |
| IRIN | **95.2** | **0.031s** | **97.9** | **0.029s** | **96.4** | **0.033s** | **96.0** | **0.036s** |
| ILOS-0.1 | 85.7 | 10.8s | 92.4 | 21.91s | 90.2 | 23.14s | 89.1 | 24.16s |
| ILOS-0.2 | 86.4 | 13.85s | 93.3 | 28.43s | 90.9 | 29.89s | 90.1 | 30.95s |
| ILOS-0.3 | 86.3 | 17.5s | 92.8 | 35.94s | 90.6 | 36.88s | 89.9 | 39.00s |
| GwFReID | 87.7 | 41.3s | 94.0 | 63.7s | 92.5 | 65.3s | 91.8 | 69.9s |

Table 1: Comparison on feedback processing time in Veri776. Here, ILOS-$0.1/0.2/0.3$ sets the weight of the distillation loss to $0.1/0.2/0.3$, respectively. In this paper, $0.2$ is the default setting.




(b) Veri776 (c) VehicleID-Small

Figure 4: Sensitivity to the size of candidate set.

the cost to update the query feature and obtain a new ranking list. For incremental learning, online training overhead is incurred to update the underlying model with the augmented positive samples by random flipping, padding and erasing. With the number of augmented samples fixed to 24, we vary the weight of distillation loss (denoted by $\lambda_d$) from $0.1$ to $0.3$ in the loss function of ILOS. As we can see from Table 1, it takes higher running time when $\lambda_d$ increases. This is because the model training becomes more "query-specific" and its online training requires more epochs to be converged. Thus, ILOS-$0.2$ is more accurate than ILOS-$0.1$ but incurs higher processing time. However, when $\lambda_d$ continues to rise, we also find that the mAP does not increase monotonically, probably because the information from the feedback signal are not sufficient to provide reliable clues to guide the optimization directions of model parameters. In contrast, IRIN is much faster and more accurate than incremental learning. It takes less than $0.04$s to handle a feedback signal and can easily support real-time user interaction, whereas ILOS requires users to wait more than $10$ seconds for each interaction. The running time of GwFReID is even worse, around $3$-$4$ times higher than that of ILOS-0.1. The reason is that GwFReID maintains a larger exemplar set than ILOS and requires higher re-training overhead.

## Experiments with Real Human Interaction

We invite $20$ postgraduates with $10$ female and $10$ male students to participate in the real human interaction experiment. Each student is required to provide iterative feedback for each query image with $5$ iterations. We observe that the quality of feedback is different for male and female students. Thus, we will report their results separately and refer them as IRIN-Male and IRIN-Female, respectively.

The comparison approaches include three modes. 1) For machine-only mode, we still use TransReID as the baseline without human interaction. 2) For human-machine cooperation mode, we compare IRIN with incremental learning method for ReID (denoted by IRIN-Male and IRIN-Female, respectively). Furthermore, we also report IRIN-Oracle to show the upper bound performance with perfect feedback signal. 3) For human-only baselines, we require the students to annotate the matching vehicles for the query images. To reduce their annotation cost, we make an assumption that the positive samples are contained in the top-200 most similar images obtained in the initial round of IRIN. This procedure generates two baselines: Male-Annotation and Female-Annotation.

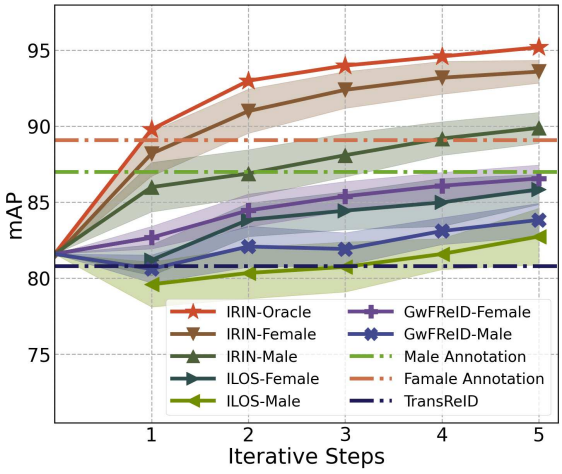



Figure 5: The mAP results with real human interaction on Veri776. Shaded areas represent the variance of mAP.

In Figure 5, we report the mean mAP results for all participants with increasing number of iterations applied on each query image. For the human-machine cooperative methods, since the quality of annotation is different among the students, we also plot the variance with shaded area. The results lead us to the following key observations. 1) The mAPs of human-only baselines are significantly higher than the machine-only baseline, indicating that human users are more capable of distinguishing the true positives from a set of similar images. 2) The results of Male-Annotation and Female-Annotation show that female students can provide more ac-

| Method | IRIN-Oracle | | | IRIN-Female | | | IRIN-Male | | |
|---|---|---|---|---|---|---|---|---|---|
| | mAP | rank-1 | rank-5 | mAP | rank-1 | rank-5 | mAP | rank-1 | rank-5 |
| IRIN | **95.2** | **99.5** | **100** | **93.2** | **98** | **99.5** | **89.9** | **96.5** | **99** |
| IRIN-max-dist | 94.5 | 99 | 99 | 92.1 | 97.5 | 99 | 88 | 96 | 98 |
| IRIN-random-sample | 93.8 | 99.5 | 99.5 | 91.8 | 97.5 | 98.5 | 87.7 | 95.5 | 98.5 |
| IRIN-triplet-loss | 92.3 | 99 | 99.5 | 90.6 | 95.5 | 98 | 85.9 | 94.5 | 97.5 |
| IRIN-w/o-HNMB | 92.7 | 99 | 99.5 | 91.3 | 97 | 98.5 | 86.8 | 96.5 | 98 |

Table 2: The ablation study of sampling and optimization strategy.

curate feedback. Consequently, when the feedback is provided to IRIN, we can see that IRIN-Female outperforms IRIN-Male. This finding is consistent with the experiment on feedback quality in Figure 3. 3) Both IRIN, ILOS and GwFReID can benefit from human feedback. Their mAPs increase steadily with the number of interactions per query. 4) Compared with ILOS and GwFReID, IRIN is more effective in leveraging the human feedback and achieves higher accuracy. 5) With sufficient number of iterations, IRIN, as a human-machine cooperation method, eventually outperforms machine-only and human-only baselines. Besides, we can observe that its variance reduces with more iterations of human interaction.

## Ablation Study

We examine four variants of IRIN for ablation study. `IRIN-random-sample` replaces the active learning based sampling strategy with random sampling to obtain the set of uncertain candidates for human verification. `IRIN-max-dist` simulates human interaction by selecting the positive sample with the maximum distance to the query image, whereas the original IRIN uses the minimum distance. `IRIN-triplet-loss` replaces contrastive loss function with traditional triplet loss. `IRIN-w/o-HNMB` removes the component of hard negative memory bank.

In Table 2, we report the results of these four IRIN variants under different settings of feedback quality. In the offline training stage, the active learning based sampling strategy for uncertain candidate selection is more effective than random sampling. With these uncertain candidates, we can see that the strategy to simulate human users to pick the sample with the minimum distance to query image is indeed superior to picking the most different candidate. Overall, the supervised contrastive loss and hard negative memory bank play more important effect in this ablation study. When either component is removed, we can observe considerable performance degradation.

## Generality to Person ReID

In the final experiment, we evaluate the generality of our framework by applying it to person ReID and conduct experiments with human interaction in **MSMT17** (Wei et al. 2018). The dataset contains $126,441$ pictures of $4,101$ pedestrians, captured by $15$ cameras. We use similar comparison methods as those in Figure 5. Since TransReID claims to achieve state-of-the-art performance in both vehicle and person ReID, we choose it as the machine baseline without human intervention. The results in Figure 6 the mAP derived from the backbone network of IRIN is inferior to TransReID ($74.6\%$ vs $79.1\%$). Nevertheless, with only one iteration of feedback, both IRIN-Male and IRIN-Female have outperformed TransReID. With more iterations, the performance advantage is further widened. When there are $5$ iterations, their mAPs are substantially higher than that of TransReID, reaching $89.3\%$ and $91.5\%$ respectively. It is also interesting to observe that IRIN-Female achieves almost the same accuracy with IRIN-Oracle when the number of iterations is set to $5$, implying that our proposed IRIN is effective in leveraging imperfect feedback.

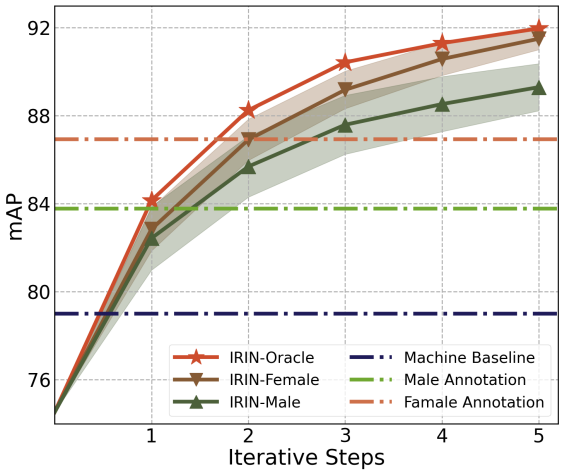



Figure 6: Experiment on person ReID.

## Conclusion

In this paper, we study vehicle ReID in a new scenario — with human in the loop to provide iterative feedback for continuous performance enhancement. We propose an Interaction ReID Network (IRIN) to effectively fuse the feedback signal and dynamically adjust the feature embedding of the query image. It can support real-time interaction and is suitable for applications with high accuracy requirement. Experimental results validate the superiority of such human-machine cooperation mode over machine-only or human-only baselines. With flawed feedback, the proposed IRIN can even outperform the state-of-the-art vehicle ReID model by a wide margin. In our future study, we will explore a more diversified set of human operators as feedback signals.

## Acknowledgments

This work is sponsored by CCF-Huawei Populus Grove Fund.

## References

Berti-Équille, L. 2019. Reinforcement Learning for Data Preparation with Active Reward Learning. In Yacoubi, S. E.; Bagnoli, F.; and Pacini, G., eds., *Internet Science - 6th International Conference, INSCI 2019, Perpignan, France, December 2-5, 2019, Proceedings*, volume 11938 of *Lecture Notes in Computer Science*, 121–132. Springer.

Brenner, R.; Priyadarshi, J.; and Itti, L. 2016. Perfect Accuracy with Human-in-the-Loop Object Detection. In *ECCV 2016*, volume 9914 of *Lecture Notes in Computer Science*, 360–374.

Ghosh, A.; Shanmugalingam, K.; and Lin, W. 2021. Relation Preserving Triplet Mining for Stabilizing the Triplet Loss in Vehicle Re-identification. *CoRR*, abs/2110.07933.

He, J.; Mao, R.; Shao, Z.; and Zhu, F. 2020. Incremental Learning in Online Scenario. In *CVPR 2020*, 13923–13932. Computer Vision Foundation / IEEE.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. TransReID: Transformer-based Object Re-Identification. In *ICCV 2021*, 14993–15002. IEEE.

Khorramshahi, P.; Kumar, A.; Peri, N.; Rambhatla, S. S.; Chen, J.; and Chellappa, R. 2019. A Dual-Path Model With Adaptive Attention for Vehicle Re-Identification. In *ICCV 2019*, 6131–6140. IEEE.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. In *NeurIPS 2020*.

Liu, H.; Tian, Y.; Wang, Y.; Pang, L.; and Huang, T. 2016a. Deep Relative Distance Learning: Tell the Difference between Similar Vehicles. In *CVPR 2016*, 2167–2175. IEEE Computer Society.

Liu, X.; Liu, W.; Mei, T.; and Ma, H. 2016b. A Deep Learning-Based Approach to Progressive Vehicle Re-identification for Urban Surveillance. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *ECCV 2016*, volume 9906 of *Lecture Notes in Computer Science*, 869–884. Springer.

Liu, Z.; Wang, J.; Gong, S.; Tao, D.; and Lu, H. 2019. Deep Reinforcement Active Learning for Human-in-the-Loop Person Re-Identification. In *ICCV 2019*, 6121–6130. IEEE.

Lou, Y.; Bai, Y.; Liu, J.; Wang, S.; and Duan, L. 2019. Embedding Adversarial Learning for Vehicle Re-Identification. *IEEE Trans. Image Process.*, 28(8): 3794–3807.

Muthuraman, K.; Reiss, F.; Xu, H.; Cutler, B.; and Eichenberger, Z. 2021. Data cleaning tools for token classification tasks. In *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, 59–61.

Pan, X.; Luo, P.; Shi, J.; and Tang, X. 2018. Two at Once: Enhancing Learning and Generalization Capacities via IBN-Net. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision - ECCV 2018 - 15th European Conference, Part IV*, volume 11208 of *Lecture Notes in Computer Science*, 484–500. Springer.

Quispe, R.; Lan, C.; Zeng, W.; and Pedrini, H. 2021. AttributeNet: Attribute enhanced vehicle re-identification. *Neurocomputing*, 465: 84–92.

Rao, Y.; Chen, G.; Lu, J.; and Zhou, J. 2021. Counterfactual Attention Learning for Fine-Grained Visual Categorization and Re-identification. *CoRR*, abs/2108.08728.

Stonebraker, M.; Bhargava, B.; Cafarella, M.; Collins, Z.; Mc-Clellan, J.; Sipser, A.; Sun, T.; Nesen, A.; Solaiman, K.; Mani, G.; et al. 2020. Surveillance video querying with a human-in-the-loop. In *Workshop on Human-In-the-Loop Data Analytics*.

Sun, Y.; Cheng, C.; Zhang, Y.; Zhang, C.; Zheng, L.; Wang, Z.; and Wei, Y. 2020. Circle Loss: A Unified Perspective of Pair Similarity Optimization. In *CVPR 2020*, 6397–6406. Computer Vision Foundation / IEEE.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person Transfer GAN to Bridge Domain Gap for Person Re-Identification. In *CVPR 2018*, 79–88. Computer Vision Foundation / IEEE Computer Society.

Wu, C.; Wu, F.; Qi, T.; Huang, Y.; and Xie, X. 2021. Fastformer: Additive Attention Can Be All You Need. *CoRR*, abs/2108.09084.

Wu, G.; and Gong, S. 2021. Generalising without Forgetting for Lifelong Person Re-Identification. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 2889–2897. AAAI Press.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 3733–3742. Computer Vision Foundation / IEEE Computer Society.

Xie, S.; Girshick, R. B.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *CVPR 2017*, 5987–5995. IEEE Computer Society.

Yan, C.; Pang, G.; Bai, X.; Zhou, J.; and Gu, L. 2020. Beyond Triplet Loss: Person Re-identification with Fine-grained Difference-aware Pairwise Loss. *CoRR*, abs/2009.10295.

Zakria; Deng, J.; Khokhar, M. S.; Aftab, M. U.; Cai, J.; Kumar, R.; and Kumar, J. 2021. Trends in Vehicle Re-identification Past, Present, and Future: A Comprehensive Review. *CoRR*, abs/2102.09744.

Zhang, D.; Li, Z.; Wang, X.; Tan, K.; and Chen, G. 2022. Towards One-Size-Fits-Many: Multi-Context Attention Network for Diversity of Entity Resolution Tasks. *IEEE Trans. Knowl. Data Eng.*, 34(12): 6018–6032.

Zhao, J.; Zhao, Y.; Li, J.; Yan, K.; and Tian, Y. 2021. Heterogeneous Relational Complement for Vehicle Re-identification. *CoRR*, abs/2109.07894.