Cloth-Changing Person Re-identification with Invariant Feature Parsing for UAVs Applications

Mingfu Xiong, Xinxin Yang, Hanmei Chen, Wael Hosny Fouad Aly, *Senior Member, IEEE*, Abdullah AlTameem, Abdul Khader Jilani Saudagar, Shahid Mumtaz, *Senior Member, IEEE*, Khan Muhammad, *Senior Member, IEEE*

***Abstract*—Recently, deep learning-based intelligent vehicle control systems have played an important role in real-time road conditions assessment applications. It relies primarily on unmanned aerial vehicles (UAVs) for specific target retrieval, especially Cloth-Changing Person Re-identification (CC-ReID) technology, to provide support for road observations and environmental monitoring. Existing CC-ReID methods mainly focus on the invariant features of the front and rear views that are independent of clothing; among them, global color enhancement is a commonly used strategy. However, this method usually reduces the chromatism between the target foreground and background, which can easily lead to the loss of features unrelated to clothing and reduce the model's performance. To solve this problem, this paper proposes a data augmentation framework with Local Invariant Feature Transformation and Clothing Adversarial Parsing (LIFTCAP) for CC-ReID. The proposed framework is equipped with a Local Invariant Feature Transition (LIFT) module and a Clothes Adversarial Parsing (CAP) module. The former aims to extract invariant features for the same person with different clothes using the local transition manners. CAP is devoted to finding adversarial associations and parsing contour differences between clothing styles. Subsequently, a feature correlation strategy is alternately implemented between the two modules to complete the optimization procedure. Extensive experiments were conducted on the public CC-ReID datasets (LTCC and PRCC), demonstrating the superiority of our proposed method over the latest methods. Furthermore, our method achieved competitive performance, particularly on a surveillance video dataset (CCVID). In addition, based on the LIFTCAP strategy, the proposed algorithm can achieve a time efficiency as low as O(n) for detecting specific targets when deployed on a UAV server (Feisi X200) for real-time road conditions assessment and monitoring applications.**

***Index Terms*—Clothes Change, Person Re-identification, Intelligent Vehicle Control, Invariant Feature, Data Enhancement, Unmanned Aerial Vehicles.**

Manuscript received October 17, 2023; Revised January 24, 2024 and March 13, 2024; Accepted April 9, 2024; Published XXXX. This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-RP23058). This paper was recommended by Associate Editor XYZ. (Corresponding authors: Hanmei Chen and Khan Muhammad).

Mingfu Xiong and Xinxin Yang are with the School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan, 430200, China (e-mails: xmf2013@whu.edu.cn, Xinxin_Yang@aliyun.com).

Hanmei Chen is with the Hubei Technology Exchange, Hubei Provincial Department of Science and Technology, Wuhan, 430064, China (e-mail: chenhanmei@51kehui.com).

Wael Hosny Fouad Aly is with the College of Engineering and Technology, American University of the Middle East, Egaila 54200, Kuwait (e-mail: drwaelaly@ieee.org).

Abdullah AlTameem and Abdul Khader Jilani Saudagar are with the Information Systems Department, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia (e-mails: altameem@imamu.edu.sa,aksaudagar@imamu.edu.sa).

Shahid Mumtaz is with the Department of Applied Informatics, Silesian University of Technology Akademicka, Gliwice, Poland, and also with the Department of Computer Science, Nottingham Trent University, Nottingham, U.K. (e-mail: dr.shahid.mumtaz@ieee.org).

Khan Muhammad is with the Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab), Department of Applied Artificial Intelligence, School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Seoul 03063, Republic of Korea (e-mail: khan.muhammad@ieee.org).

I. Introduction

RECENTLY, connected and autonomous vehicles technology based on artificial intelligence (AI) and deep learning has been widely applied in real-time road conditions monitoring, where intelligent specific target detection system equipped with the unmanned aerial vehicles (UAVs) has played a major role [1], [2]. It primarily relies on retrieval technology, particularly the support of Person Re-identification (ReID) technology [3], [4]. Most traditional ReID studies assume that people do not change clothes in a period of time [3], [5]. However, in the real process of target retrieval and criminal investigation, it is normal for the target objects to change clothes. The existing ReID methods face significant challenges and cannot be applied. Therefore, it is necessary to consider the Cloth-Changing Person Re-identification (CC-ReID) technology. Different from traditional ReID, the Cloth-Changing Person Re-identification (CC-ReID) aims to judge the persons who wear different clothes in varying time-frames, scenes, or viewpoints belonging to the same identify or not, which fulfills greater challenges and difficulty [6]. Because of the traits of CC-ReID, it has been widely used in criminal investigation [4], [7], target retrieval [3], [8], and complex scenario analysis tasks [9], which have also attracted widespread attention in recent years.

It is well known that people can recognize familiar friends at a glance even if they are wearing clothes that have never been seen before [6]. This is mainly because the person's brain can quickly extract features (parsing contour, grayscale, etc.) unrelated to clothes on friends' bodies and recognize them quickly. Therefore, the extraction of invariant features that are unrelated to clothing is currently the focus of CC-ReID research. Existing CC-ReID methods typically attempt to learn the contours [10], skeletons [11], 3D shape feature [12], or the multi-modal information fusion for each individual [13], [14]. Although these methods have achieved impressive results, there is still room for further improvement, owing to the unicity of invariant features and the complexity of model construction. In addition, to increase the diversity of

training data, data augmentation methods have been proposed for CC-ReID [6]. These methods mainly focus on the invariant features of the front and rear views that are independent of clothing. Among them, global color enhancement methods are commonly used strategies [6]. However, these methods usually reduce the chromatism between the target foreground and background, which can easily lead to the loss of features unrelated to clothing (parsing contour, grayscale, etc.) and reduce the performance of the model for ReID.

To solve this problem, we propose a data augmentation framework with Local Invariant Feature Transformation and Clothing Adversarial Parsing (LIFTCAP) for CC-ReID. The proposed framework is equipped with a Local Invariant Feature Transition (LIFT) module and a Clothes Adversarial Parsing (CAP) module. Specifically, to extract invariant attributes for the same person with different clothes, the LIFT module is used to perform the local transition manners (Random Regions Erasing, Regions Color Changing. etc.) to obtain inherently invariant features for the same. The robustness of the model can be improved by incorporating the previous transition manners. In addition, the CAP is devoted to finding adversarial associations and parsing contour differences between clothing styles, which attempts to search for the correlation characteristics among different clothes. Subsequently, a feature correlation strategy is alternately implemented between the two modules to complete the optimization procedure.

Our proposed CC-ReID algorithm has a high rate of retrieving special targets, which could reach a time efficiency as low as $O(n)$. It can be applied to mobile devices for environmentally sustainable monitoring, such as the unmanned aerial vehicles (UAVs) server (Feisi X200) for real-time road conditions assessment applications [9], [15]. In addition, our LIFTCAP CC-ReID algorithm can be deployed in smart city alarms and connected and autonomous vehicle systems (http://www.autolabor.cn/pro/detail/4) for urban traffic safety monitoring. Specifically, in terms of urban traffic management, autonomous vehicle systems can help urban transportation departments monitor road conditions in real-time, release road condition information promptly, remind drivers to choose suitable routes, and avoid congestion and traffic accidents.

Extensive experiments are conducted to evaluate the performance of the proposed method. The results show that the proposed method performs better than the existing invariant features learning and multi-modal fusion methods. In addition, the local random regions invariant feature learning and Clothes Adversarial Parsing manner significantly improve the accuracy of CC-ReID compared with the global color enhancement operation on the LTCC [11], PRCC [12] and CCVID [16] datasets, respectively.

The main contributions of this study are as follows:

- This study has proposed a Local Invariant Feature Transformation and Clothing Adversarial Parsing (LIFTCAP) framework, which includes a Local Invariant Feature Transition (LIFT) module and a Clothes Adversarial Parsing (CAP) module for the CC-ReID problem.
- The LIFT module is used to extract the inherent invariant information to retain the features unrelated to clothing for the same, and the CAP module is devoted to finding the adversarial associations between clothing styles. The proposed modules improve the robustness of the framework and perform well in the CC-ReID task.
- This study also proposed a solution to deploy the CC-ReID algorithms into practical road conditions assessment applications, bridging the gap between theories and practice, which is also suitable to the other common vision tasks.
- Extensive experiments are conducted on the public CC-ReID datasets (LTCC, PRCC, and CCVID), which are used to show the competitive performance of the proposed method. It also verifies that the local invariant feature exploration manner performs better than the global color enhancement ones.

II. Related Work

A. Cloth-Changing

ReID Methods As described above, the core of CC-ReID is devoted to extracting features that are not related to clothing, (such as face, gait, and appearance). In addition to providing a new dataset, PRCC [12] extracted the contour sketch of a person image for cross-clothes ReID to moderate clothing change. LTCC [11] designed a method to extract the soft-biometrics feature to eliminate clothing appearance features that focused on body shape information. CCVID [16] proposed a Clothes-based Adversarial Loss (CAL) to learn the irrelevant feature for clothes from RGB images, which penalized the discriminative power of ReID. In addition, AIM [17] was proposed to alleviate clothing bias using a dual-branch model to mine discriminative ID cues for CC-ReID.

B. Data Augmentation Methods

Recently, data augmentation-based techniques (such as random cropping and flipping) have been employed to address the CC-ReID issue. The random erase algorithm proposed by [18] attempts to simulate the occlusion frequently encountered in reality by randomly erasing a part of the image to solve the occlusion problems for CC-ReID. Zheng et al. [6] utilized generative adversarial networks to replace the clothes of one person with those of other people to generate more diverse data and improve the generalization ability of the model. Gong et al. [19] proposed a local transformation attack (LTA) and the joint adversarial defense (JAD) method to enhance the contour or color information, which considered the local homomorphic transformation for the CC-ReID problem. Jia et al. [20] reinforced person-unrelated feature learning by designing powerful complementary data augmentation strategies that included both positive and negative data augmentation schemes. CCAF [21] attempted to expand the cloth-changing data via personal features rather than the original images, which added the diversity of clothes color and texture variations for feature distribution.

C. UAVs-based for ReID Methods

With the development of UAV technology, person ReID technology from its perspective has gradually become a popular research topic. Zhang et al. [22] proposed an airborne person ReID dataset that covered real UAV surveillance scenarios.

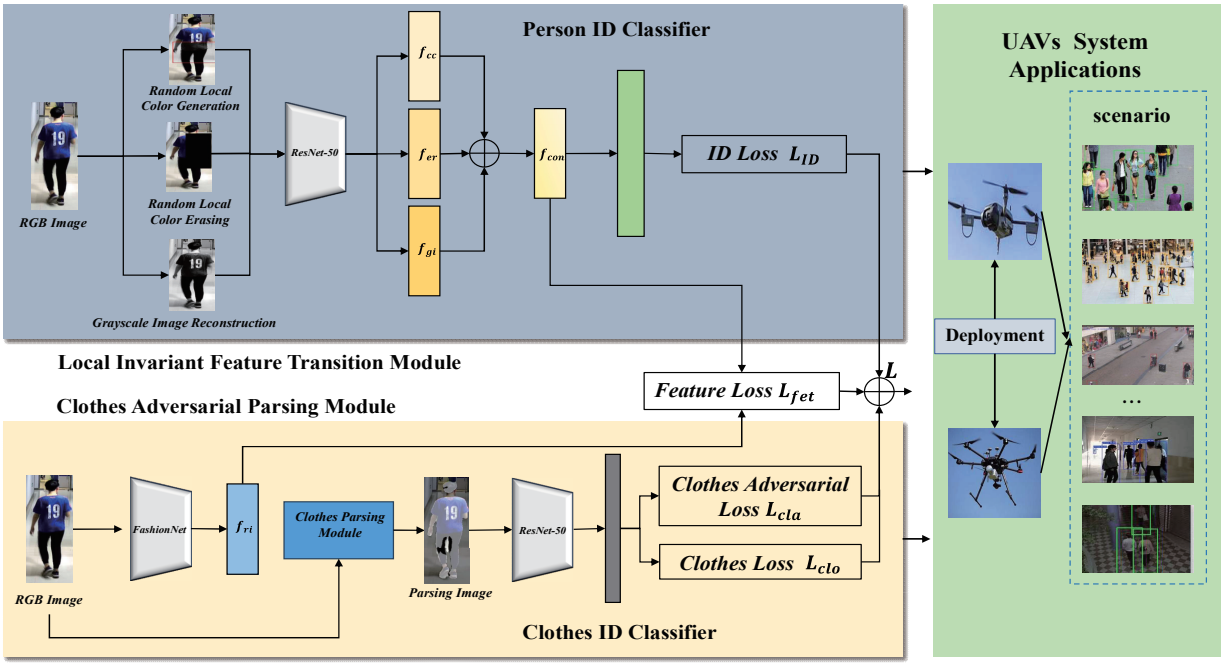

Fig. 1: The framework of the proposed LIFTCAP method with the LIFT module, CAP module, and for UAVs' application.

They utilized subspace pooling of convolution operations for person representation. Zhang et al. [23] utilized an edge-based Federated Learning Framework for Person Re-identification in UAV Delivery Service, which efficiently located target receivers. Wang et al. [24] constructed a dataset for vehicle re-identification (ReID) that distinguished a particular vehicle from others from a UAV perspective. Dong et al. [25] presented a method equipped with federated learning for person re-identification using Blockchain-Integrated Smart UAV Delivery Systems that analyzed the system's resilience under various security attacks.

Although existing CC-ReID methods have received extensive research from scholars, our method focuses more on the local Invariant Feature learning and Clothing Adversarial Parsing strategy, which is more discriminative and effective for CC-ReID performance.

III. The Proposed Method

A. Local Invariant Feature Transition Module

To increase the diversity and chromatism intensity of the background and foreground in real surveillance scenes, we propose a Local Invariant Feature Transition (LIFT) module, which enhances the generalization ability of the model to mine features unrelated to clothing. Three strategies are executed on the original RGB images, as shown in Fig. 1. First, a random local color generation operation was used to increase the diversity of different clothing colors. Then, the robustness of the transition module is improved via random local color erasing, which also improves the chromatism intensity between the person and the background. Finally, the grayscale image reconstruction operation ensured the integrity of the original target. Next, each operation is discussed in detail.

Random Local Color Generation: To extract features unrelated to clothing, we need to adapt to as many different appearances of clothes as possible. Therefore, a random local color generation operation is performed on the entire batch of images with a probability during model training. The initialization of the Random Local Color Generation operation is described by Eq. (1):

$$P_c^0 = P, \tag{1}$$

where P denotes the original image and P_c^0 is the image without any transformation. Setting the probability of random color generation as Pro, and the remaining probability is $1 - Pro$. Then, we set a certain probability for random sampling. If the sampling result is less than Pro, no random color generation operation is performed, and the original RGB image is returned. Otherwise, the operation is performed. The iterative optimization of Local Color Generation is defined by Eq. (2) and (3).

$$P_c^n = ColorGeneration(P), \tag{2}$$

$$P_c^{n+1} = \Psi_P^{\varepsilon}(P_c^n + \alpha \cdot sign(grad^{n+1})), \tag{3}$$

where $ColorGeneration(\cdot)$ denotes the random color generation. The specific execution operation is expressed by Eq. (3). P_c^n represents the random color generated in the n-th iteration. Ψ_P^{ε} is a random area selection operation that ensures that there is no noticeable noise during random color generation and that the noise is maintained within a certain range. α is a weight parameter that balances the relative gradient and importance of color generation. $sign(\cdot)$ is used to preserve the directional information of the gradient, and $grad^{n+1}$ is the gradient of the loss function for the random color generation P_c^n parameter, which shows the direction of the loss function changes for the color transformation.

Random Local Color Erasing: Color erasing is a data augmentation technique, which randomly selects a rectangular area in a person's image and sets the pixel values of that area to random values. In this study, we exploited a random local color-erasing operation to improve the chromatism intensity between the s and the background. This causes certain parts of the image to be erased, thereby simulating occlusion or damage. The purpose was to encourage the model to learn more robust features, enabling it to better handle incomplete or damaged images, thereby improving the generalization ability and performance of the model. The specific implementation process is shown in Eq. (4).

$$I_{\text{erased}}(x, y) = I(x, y) \cdot \text{mask}(x, y), \tag{4}$$

where x and y are the horizontal and vertical coordinates, respectively, of the pixels in the image. $I(x, y)$ denotes the pixel value of the original image. $I_{\text{erased}}(x, y)$ represents the pixel value after the erasing operation, and mask(x, y) is a binary operation of the same size as that of the original image, where the pixel values in the erased area are set to 0 and the other areas are set to 1. Its definition is given by Eq. (5).

$$\text{mask}(x, y) = \begin{cases} 0 & \text{with probability } r \\ 1 & \text{otherwise} \end{cases}, \tag{5}$$

where r denotes the probability of being erased. In ReID tasks, this is usually set to 0.5 [18].

Grayscale Image Reconstruction: To eliminate color interference while preserving the integrity of the original target, grayscale image reconstruction is a worthwhile choice. Firstly, we set a weight pr, which means the probability of converting the original color image into a grayscale image (p_r is set as 0.05 in this work). Then, the final reconstructed grayscale image is composed of the original color image and the to-be-converted grayscale image, which is combined with a certain probability (p_r). This strategy allows us to retain a series of color features when reconstructing the grayscale image, rather than choosing whether or not to perform a grayscale transformation completely at random. This procedure is described as Eq. (6).

$$P_o = (1 - p_r) \cdot P_i + p_r \cdot P_{gray}, \tag{6}$$

where P_i denotes the original color image. P_{gray} denotes the to-be-converted grayscale image. P_o denotes the reconstructed grayscale image.

After the above three operations, using ResNet-50 [26] as the backbone network, we obtained three types of features (color generation, color erasing, and grayscale image features), which are denoted as f_{cc}, f_{er} and f_{gi}, respectively. Next, we fuse these features to obtain a feature f_{con} that integrates multiple local transition modalities via the linear combination manner, which is shown in Eq. (7). The ID loss function was used to train this process, as shown in Eq. (8) and Eq. (9).

$$f_{con} = f_{cc} + f_{er} + f_{gi}, \tag{7}$$

$$L_{ID} = -\frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N} y_{i,j} \log \left(\frac{e^{s \cdot d_{i,j}}}{\sum_{k=1}^{N} e^{s \cdot d_{i,k}}} \right), \tag{8}$$

$$d_{i,j} = cos(f_{con}^{i}, f_{con}^{j}), \tag{9}$$

where N is the batch size and M is the total person ID. i and j represent the indices of the samples. $y_{i,j}$ is an indicator function that takes the value 1 when the identities of samples i and j are the same and 0 otherwise. $d_{i,j}$ and $d_{i,k}$ denote the cosine similarity between samples i and j (k). s is an adjustable scale parameter that controls the measurement range of the feature distance.

B. Clothes Adversarial Parsing Module

Generally, although the styles of popular clothes that we wear differ, their basic structure and contours remain somewhat similar. If these similar associations are found, identifying cross-clothing associations would be very effective. In this section, we propose a Clothes Adversarial Parsing (CAP) module equipped with two branches to determine the adversarial associations and parsing contour differences between clothes styles. The first branch is used to extract clothing features with FashionNet [27] from the original RGB image, represented as f_{ri}. It can also be seen as an original person feature in which clothes and s are integrated. Then, the person image parsing feature is represented by f_{pi}. Compared with the fused feature f_{con}, it is another form of feature expression for the same. Therefore, a new optimization loss called the Adversarial Feature Error (AFE) loss L_{fet} is proposed to train the procedure and is described by Eq. (10).

$$L_{fet} = \lambda \cdot L_{feat}(f_{con}, f_{ri}) + (1 - \lambda) \cdot L_{feat}(f_{con}, f_{pi}), \tag{10}$$

where $L_{feat}(\dot{})$ is the feature consistency loss for measuring f_{con} and f_{ri} and is represented by Eq. (11). λ is a hyperparameter that controls the loss weights of the two components and is set 0.5 [28].

$$L_{feat}(\dot{}) = \frac{1}{N} \sum_{i=1}^{n} (f_{con}^{i} - f_{ri}^{i})^2, \tag{11}$$

where N is the batch size, f_{con}^{i} and f_{ri}^{i} represent the i-th feature sample from the fused and clothing features, respectively.

The second branch is used to obtain the corresponding parsing feature from the person's body, which also processes the original RGB image more deeply. This step was used by the person parsing network [29] to obtain the parsing contour information for each person, which was fed into RestNet-50 and predicted by the Clothes ID Classifier for person classification. The entire process involves two loss functions, L_{clo} and L_{cla}, inspired by [16] and represented by Eqs. (12) and (13).

$$L_{clo} = \sum_{i=1}^{N} \left((f_i \cdot \varphi_{y_i^{clo}} - \log \left(\sum_{j=1}^{N_{clo}} e^{f_i \cdot \varphi_j / \tau} \right) \right), \tag{12}$$

where N is the batch size, f_i is the i-th feature, and $\varphi_{y_i^{clo}}$ is the parameter of the real clothing label. N_{clo} denotes the total number of clothing categories. φ_j denotes the parameter vector of the j-th clothing category. τ is a temperature parameter that controls the degree of smoothing of clothing category

$$L_{cla} = -\sum_{i=1}^{N}\sum_{clo=1}^{N_{clo}} q(clo)(f_i \cdot \varphi_{clo}/\tau - \log(e^{(f_i \cdot \varphi_{clo}/\tau)} + \sum_{j \in S_i^-} e^{(f_i \cdot \varphi_{clo}/\tau)})), \tag{13}$$

distribution. f_i denotes the feature vector of the i-th individual. φ_{clo} is the feature vector and clo is used to refer to the category of clothing, which is indicated by the category of the first piece of clothing, S_i^+ (S_i^-) is the set of clothing classes with the same identity (different identities) as f_i. K is the number of classes in S_i^+, and q_{clo} is the weight function of the cross-entropy loss, which adjusts the loss contribution according to whether the classes belong to the set of clothing classes of the same identity, that is, the positive classes with the same clothes ($clo = y_i^{clo}$) and the positive classes with different clothes ($clo \neq y_i^{clo}$ and $clo \in S_i^+$), that is $1/K$. This process is described by Eq. (14).

$$q(clo) = \begin{cases} \frac{K-\epsilon(K-1)}{K} & clo = y_i^{clo} \\ \frac{\epsilon}{K} & clo \neq y_i^{clo} \text{ and } clo \in S_i^+, \\ 0 & clo \in S_i^- \end{cases} \tag{14}$$

where ϵ is the balance weight to adjust the model's sensitivity to changes in clothing. In this setup, $q(clo)$ is the loss function.

C. The Optimization and Loss Functions

An end-to-end optimization method is used to train the proposed LIFTCAP framework with four types of loss functions. Subsequently, the overall loss function L is expressed as Eq. (15).

$$L = L_{ID} + L_{fet} + L_{cla} + L_{clo}, \tag{15}$$

where L_{ID}, L_{fet}, L_{cla},L_{clo} represent the person identity loss, adversarial feature error (AFE) loss, clothes adversarial loss, and clothes loss, respectively.

D. Discussion with Previous Methods

As mentioned in Section II, to excavate invariant features unrelated to clothing, existing methods such as CCVID [16] and FRGS [19] have also proposed related algorithms and strategies to address this problem. However, there are essential differences between our method for existing methods, which are summarized as follows:

- Firstly, CCVID [16] learn the invariant features just with the RGB modality for clothing change. It can be easily influenced by the color difference between foreground and background. In our work, we have extended and adapted these methods, which obtain inherent invariant features via the local transition manner for the same person with different clothes.
- Secondly, FRGS [19] mainly exploited the local color variation transformation attack and adversarial defense methods for conventional ReID (Clothing remains unchanged during a period). We have proposed the Adversarial Feature Error (AFE) for CC-ReID.
- Thirdly, we do not just consider the clothes styles and categories but try to mine the local invariant features and their association. In addition, we also propose a solution to deploy the CC-ReID algorithms into UAV applications, bridging the gap between theories and practice, which is also suitable to the other common vision tasks.

In summary, there are essential differences between the proposed method and existing methods in terms of the problem object, feature discovery, and optimization procedure.

IV. Experiments and Results

A. Datasets and Evaluation Protocol

Datasets: The PRCC [12] dataset consists of 33,698 images from 221 identities. Within the PRCC dataset, 17,896 images of 150 people were designated as the train set, 5,002 images of 150 s as the validation set, and 3,384 images of 71 persons as the gallery set. LTCC [11] comprises two subsets: cloth-change (91 individuals, 14,783 images) and cloth-consistent (61 individuals, 2,336 images), with 77 identities contributing to the 9,576-image train set, 75 identities forming the 7,050-image test set, and another 75 identities comprising the 493-image query set. The CCVID [16] dataset contains 226 identities and 2,856 sequences. Among these, 75 identities and 948 sequences were allocated to the train set and 151 identities were designated for testing. 1,074 sequences formed the gallery set within the test set, whereas the remaining 834 sequences served as the query set.

Evaluation Protocol and Settings: Like traditional ReID, the cumulative match curve (CMC) [30] and mean average precision (mAP) [31] are used to measure the performance for CC-ReID.To evaluate the model's performance, the "Normal, CC, and CU" denoted as the normal, cloth-changing, and cloth-unchanging settings respectively.

B. Implementation Details

We selected the dual-card RTX 2080 Ti GPU with CUDA 11.1 for PRCC and LTCC datasets, and RTX A5000 GPU for CCVID datasets, respectively. The programming environment is based on Python 3.8.10, and it runs on the Ubuntu 18.04 operating system. ResNet-50 was used as the backbone network, which has been removed from its last down-sampling part to better meet the needs of this study. During processing, the input images were resized to 384×192. In the model training stage, the batch size was set to 64, and each batch contained 8 samples and their corresponding 8 images. We chose the Adam optimizer [32] to train the model for 60 epochs and introduced L_{cla} after the 25th epoch to further optimize the training process. The initial learning rate was set to 3.5×10^{-4} and was reduced to one-tenth of the original after every 20 epochs. Following the suggestion of Gu and Gong [16], [19], the parameter τ was set to $1/16$ in our method. In addition, we stipulate that each input image has a probability of 0.2 (i.e., $P_{ro} = 0.2$) for 2 and 3 specific operations. For the CCVID

TABLE I: Comparison with state-of-the-art methods on LTCC dataset.

Methods	Venue	Normal				CC			
		top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP
RestNet-50 [26]	CVPR′16	58.82	-	-	25.98	20.08	-	-	9.02
HACNN [32]	CVPR′18	60.24	-	-	26.71	21.59	-	-	9.25
Face [33]	CVPR′18	60.44	-	-	25.42	22.10	-	-	9.44
PCB [34]	ECCV′18	65.11	-	-	30.60	23.52	-	-	10.03
MGN [35]	MM′18	70.59	79.31	82.76	35.10	29.85	45.15	51.02	13.87
OSNet [36]	ICCV′19	66.07	-	-	31.18	23.43	-	-	10.56
MuDeep [37]	TPAMI′19	61.86	-	-	27.52	23.53	-	-	10.23
BOT [38]	CVPR′19	72.21	81.74	83.60	34.75	28.82	44.64	52.30	12.67
CESD [11]	TPAMI′20	71.39	-	-	34.31	26.15	-	-	12.40
FSAM [39]	CVPR′21	73.20	-	-	35.40	38.50	-	-	16.20
GI-ReID [13]	CVPR′22	73.59	-	-	36.07	28.86	-	-	14.19
Pos-Neg [20]	ACM'22	75.66	83.57	86.41	37.00	36.22	50.77	56.12	14.43
Baseline [16]	CVPR′22	73.40	82.60	85.40	39.20	38.00	51.80	56.60	17.10
AIM [17]	CVPR′23	76.30	-	-	41.10	40.60	-	-	19.10
LIFTCAP	Ours	74.60	81.30	84.60	39.70	37.00	53.10	58.40	17.90

dataset, each frame was resized to 256×128 pixels, batch size was set to 32, and each batch contained 8 characters and 4 video clips. The model was also trained for 150 epochs using the Adam optimizer, and L_{cla} was introduced after the 50th epoch. The learning rate had an initial value of 3.5×10^{-4}, which was reduced to one-tenth of the original after every 40 epochs.

C. Comparison to the State-of-the-art Methods

The proposed method was compared with the latest algorithms for solving the CC-ReID task. The mAP (%) and Top-1 accuracy of CMC (%) were evaluated over the three public datasets mentioned previously. The obtained results are listed in Tables I, II and III, respectively.

Results on LTCC dataset: In Table I, we summarize the performance of the proposed LIFTCAP and that obtained by state-of- the-art competitors on the LTCC dataset. Fourteen methods were included in the comparison benchmark: RestNet-50 [26], HACNN [32], Face [33], PCB [34], MGN [35], OSNet [36], MuDeep [37], BOT [38], CESD [11], FSAM [39], GI-ReID [13], Pos-Neg [20], baseline [16], AIM [17]. Table I shows that the proposed method achieved competitive results.

Results on PRCC dataset: We have compared our method with the latest alternatives for the CC-ReID task on the PRCC dataset. The results are show in Table II and the new methods are included as: HACNN [32], PCB [34], IANet [40], SPT+ASE [12], GI-ReID [13], RCSANet [41], FSAM [39], baseline [16], DCR-ReID [42]. The results showed that the proposed method exhibited a significant improvement relative to the baseline. Although top-1 and mAP have similar performances in the CU setting, there is a significant improvement in the CC setting. Specifically, top-1 improved from 52.2% to 56.7% and mAP improved from 54.2% to 57.7% under the CC setting.

TABLE II: Comparison with state-of-the-art methods on PRCC dataset.

Methods	Venue	CU		CC	
		top-1	mAP	top-1	mAP
HACNN [32]	CVPR′18	82.5	-	21.8	-
PCB [34]	ECCV′18	99.8	97.0	41.8	38.7
IANet [40]	CVPR′19	99.4	98.3	46.3	45.9
SPT+ASE [12]	TPAMI′21	64.2	-	34.4	-
GI-ReID [13]	CVPR′22	80.0	-	33.3	-
RCSANet [41]	ICCV′21	100	97.2	50.2	48.6
FSAM [39]	CVPR′21	98.8	-	54.5	-
Baseline [16]	CVPR′22	100	99.7	52.2	54.2
DCR-ReID [42]	TCSVT′23	100	99.7	57.2	57.4
LIFTCAP	Ours	100	99.8	54.3	55.6

Results on CCVID dataset: The CCVID is a video dataset and we have also listed some previous methods to compare with our method. The results are listed in Table III. Eight methods were included in the comparison dataset: I3D [43], Non-Local [44], GaitNet [45], GaitSet [10], AP3D [3], TCLNet [4], baseline [16], and DCR-ReID [42]. As shown in Table III, the proposed method outperformed the baseline in all metrics. Specifically, under normal settings, our method achieved top-1 and mAP of 86.3% and 84.1%, respectively, which were 3.3% higher than the baseline. Under the CC setting, the baseline had top-1 and mAP of 82.6% and 79.6%, respectively, whereas our method achieved 85.7% and 83%, respectively. Our method exhibits a significant advantage over image datasets. In addition, it surpassed the DCR-ReID method by approximately 2% in each metric. These results indicate that our method possesses high accuracy and stability in CC-ReID tasks, suggesting a performance improvement over the baseline method.

D. Ablation Studies

We conducted ablation experiments to verify the effectiveness of the proposed method. Specifically, for the CC-

TABLE III: Comparison with state-of-the-art methods on CCVID dataset.

Methods	Venue	Normal		CC	
		top-1	mAP	top-1	mAP
I3D [43]	CVPR′17	79.7	76.9	78.5	75.3
Non-Local [44]	CVPR′18	80.7	78.0	79.3	76.2
GaitNet [45]	CVPR′19	62.6	56.5	57.7	49.0
GaitSet [10]	AAAI′19	81.9	73.2	71.0	62.1
AP3D [3]	ECCV′20	80.9	79.2	80.1	77.7
TCLNet [4]	ECCV′20	81.4	77.9	80.7	75.9
Baseline [16]	CVPR′22	83.0	80.8	82.6	79.6
DCR-ReID [42]	TCSVT′23	84.7	82.7	83.6	81.4
LIFTCAP	Ours	86.3	84.1	85.7	83.0

TABLE IV: Probability of LIFTCAP on LTCC.

Probability	CU		CC		Normal	
	top-1	mAP	top-1	mAP	top-1	mAP
Baseline [16]	80.6	64.4	38.0	17.1	73.4	39.2
Pro=0.2	81.1	64.4	39.3	17.0	74.2	39.3
Pro=0.4	81.3	64.4	37.5	18.0	73.6	39.7
Pro=0.6	80.3	65.3	36.0	17.3	73.4	40.0
Pro=0.8	79.9	63.4	38.5	17.3	73.8	39.1

ReID task, we need to make the model learn to match and recognize the target person on different datasets. To enhance the robustness and generalization ability of our framework, the proposed LIFTCAP framework has set different parameters on the public datasets, as listed in Table IV. The values of Pro were set to 0.2, 0.4, 0.6, and 0.8 for the ablation experiments. The results for the LTCC dataset showed that the evaluation metrics were relatively better when set to 0.2 and 0.4. The same strategy was applied to the CCVID and PRCC datasets, as listed in Table VII.

In addition, based on the results obtained in IV, we set Pro to 0.2 and conducted ablation experiments on the PRCC dataset in the 2080 graphics card configuration with the following settings:

1) **Mask**: The original RGB image is transformed into a mask image by the existing body parsing network [29].
2) **Mask+Color Generation**: Augmentation of the data with one LIFTCAP on top of 1).
3) **RGB+Mask**: Multimodal input. The original RGB image and mask image are input to the backbone network separately, and the two features obtained are stitched together before the fully connected layer.
4) **RGB+Mask+Color Generation** : Use the Color Generation data augmentation once based on setting 3).
5) **RGB+Mask+2D** : Using random shift data augmentation [46] The images based on 3) .
6) **RGB+Mask+Color Generation+2D**: Apply random 2D translation data augmentation on top of setting 4).

As shown in Table V, the random color generation operation with only RGB modality achieved the best performance compared to the multi-modality case, and the random local color erasing operation was much lower than the previous one. This demonstrates the superiority of the proposed method for the color generation operation of a single modality.

Similarly, we also fixed Pro to 0.2 and performed ablation experiments on the LTCC dataset with 2080 cards. The details

TABLE V: Ablation Studies of LIFTCAP on PRCC. Mask stands for the image obtained by person body analysis; RGB+Mask stands for the multi-modal; 2D stands for the random 2D translation.

Methods	CU		CC	
	top-1	mAP	top-1	mAP
Baseline [16]	100	99.7	52.2	54.2
Mask [29]	99.9	99.0	46.1	47.6
Mask+Color Generation	99.6	97.4	34.7	35.9
RGB+Mask	63.2	29.0	28.0	11.9
RGB+Mask+Color Generation	64.4	29.6	28.0	12.0
RGB+Mask+Color Generation+2D [46]	56.4	24.8	25.2	11.1
RGB+Mask+2D	56.2	24.6	25.3	11.0
LIFTCAP	100	99.8	54.1	54.8

TABLE VI: Ablation Studies of LIFTCAP on LTCC(A, B) where A represents the weight of mutual information and B represents the weight of KL dispersion.

Methods	CU		CC		Normal	
	top-1	mAp	top-1	mAP	top-1	mAp
Baseline [16]	80.6	64.4	38.0	17.1	73.4	39.2
Mask [29]	77.5	61.6	27.0	14.1	69.2	36.8
Mask+Color Generation	64.3	47.7	21.9	11.3	58.2	28.8
Mask+Color Generation+(0.5,0.1)	78.9	60.6	30.9	14.9	71.6	37.0
Mask+Color Generation+(1.0,1.0)	69.5	51.4	23.5	11.6	62.1	30.9
FRGS [19]+(0.5,0.1)	75.3	56.1	28.1	12.8	67.1	33.5
Color Generation+FRGS	76.3	57.2	31.4	14.3	69.4	34.7
LIFTCAP+(0.5,0.1)	79.6	63.5	36.2	16.9	72.0	39.0
LIFTCAP	81.1	64.4	39.3	17.0	74.2	39.3

are as follows:

1) **Mask+LIFTCAP+(0.5,0.1)**: The obtained mask images are augmented with LIFTCAP data once, and in addition, KL divergence and mutual information [16] are used to guide the model learning, and their weights are set to 0.5 and 0.1, respectively (these values are determined using the control variable method, with the KL divergence weight set to 0.5 by random selection, and then the mutual information weight is determined to be 0.1 by ablation experiments. details about the different weights leading to different ReID effects renderings can be found in Fig. 2 and Fig. 3).
2) **FRGS [19]+(0.5,0.1)**: Employing the data augmentation approach proposed in [19] (this method combines the RGB channels of visible images, grayscale images, and sketch images to form a new image), with KL divergence and mutual information weights set to 0.5 and 0.1, respectively.

Table VI lists the effects of using two data augmentation methods with different KL divergence and mutual information weights on the LTCC dataset. Specifically, we first replaced the original RGB images with mask images using an existing person parsing network. We then randomly initialized the weight of the mutual information to 0.5 and controlled the weight of the mutual information to be constant. As shown in Fig. 2, when the weight of the mutual information is 0.5, the best person ReID effect on the LTCC dataset is achieved, and the weight of the KL scatter is 0.1. Then, we set the weight of the KL divergence to 0.1 and increased the weight of mutual information from 0 to 1 by 0.1. Finally, as illustrated in Fig. 3, we can see that the weight of the mutual information should be taken as 0.1. The best combination of the KL divergence and mutual information was set as (0.5,0.1). As shown in Table VI,

TABLE VII: The ablation studies on CCVID, LTCC, and PRCC. LIFTCAP (F/S,a/a+b) represents the method of Local Invariant Feature Transformation and Clothing Adversarial Parsing. F represents the usage of Random Local Color Generation once, whereas S represents the usage of Random Local Color Generation twice. In LIFTCA P(F/S,a+b), it denotes the probability value in the first Random Local Color Generation method, and b denotes the probability value in the second Random Local Color Generation method.

Methods	**CCVID**				**LTCC**				**PRCC**			
	Normal		CC		Normal		CC		CU		CC	
	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
Baseline [16]	83.0	80.8	82.6	79.6	73.4	39.2	38.0	17.1	100	99.7	52.2	54.2
LIFTCAP(F,0.4)	84.3	82.4	83.2	81.1	73.6	39.7	37.5	18.0	100	99.8	52.3	54.9
LIFTCAP(F,0.2)	85.5	83.5	85.0	82.2	74.2	39.3	39.3	17.0	100	99.8	53.8	56.0
LIFTCAP(S,0.4+0.4)	85.9	84.8	85.1	83.5	72.2	39.3	35.2	17.1	100	99.8	56.7	57.7
LIFTCAP(S,0.2+0.2)	87.2	85.0	86.3	83.8	72.6	40.0	35.7	17.9	100	99.9	53.2	55.0
LIFTCAP(S,0.2+0.4)	86.3	84.1	85.7	83.0	74.6	39.7	37.0	17.9	100	99.8	54.3	55.6

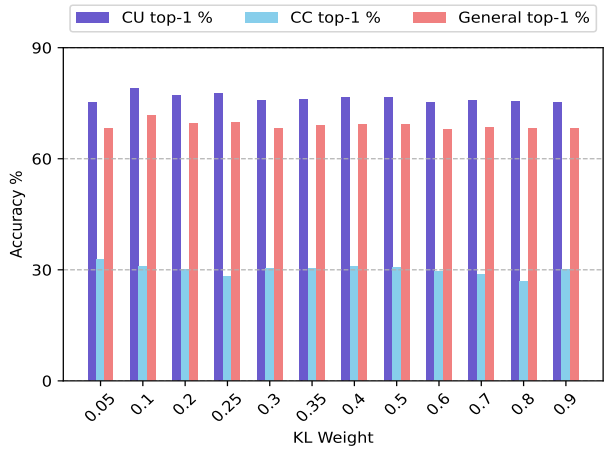

Fig. 2: Top-1 accuracy variation with different KL Weights.

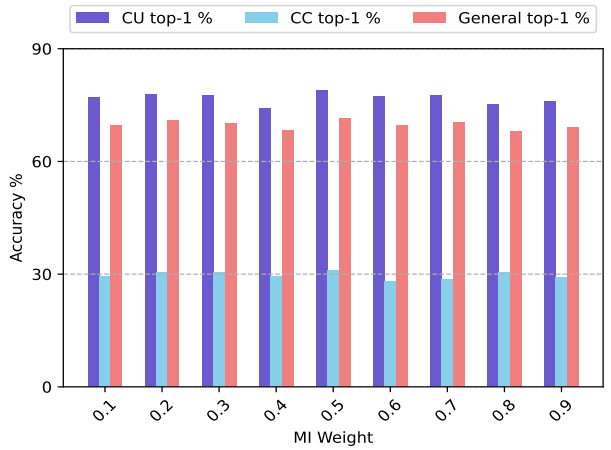

Fig. 3: Top-1 accuracy variation with different MI weights.

our method does not require additional losses to constrain the entire model and is more suitable for the CC-ReID task than the other data augmentation methods.

According to Table IV, comparing the probabilities of LIFTCAP, we determined that the probability of 0.2 or 0.4 would yield good results on the LTCC dataset. Therefore, we conducted comparative experiments on two image datasets and one video dataset using these two probabilities, as listed in Table IV. Based on Table IV, the probability of LIFTCAP performing well is 0.2 or 0.4. With the LIFTCAP (S, 0.2+0.2) configuration, the CCVID dataset showed a significant improvement compared to the baseline. However, the LTCC and PRCC datasets exhibited slightly inferior performance compared to the baseline. This can be attributed to the fact that the LIFTCAP (S, 0.2+0.2) configuration is more suitable for video datasets and has less impact on image datasets. Therefore, the LIFTCAP (S, 0.2+0.4) configuration was applied in our method. In this setting, all three datasets demonstrate a noticeable improvement. Moreover, the video dataset exhibited a higher improvement magnitude, ranging from 3% to 4%, compared with the image datasets. Consequently, we determined the optimal probability combination for LIFTCAP to be (0.2, 0.4).

TABLE VIII: Time efficiency for our algorithm on the public CC-ReID datasets (LTCC, PRCC, and CCVID).”h:m:s” denotes the time (hour, minutes, and second)

h:m:s	LTCC		PRCC		CCVID	
	Baseline	Ours	Baseline	Ours	Baseline	Ours
Train	1:29:05	1:28:49	2:46:36	2:39:14	1:52:30	1:52:29
Total	1:37:08	1:36:59	2:58:50	2:50:53	3:13:40	3:13:42

In addition, we also analyzed the time efficiency of the proposed algorithm, which can be seen in Tab. VIII. The table shows that our method achieved considerable time efficiency compared to the baseline on LTCC, PRCC, and CCVID datasets. Overall, our LIFTCAP framework reduces the total training time by introducing a Local Invariant in the feature extraction phase and fine-tuning the clothing attributes in the Adversarial Parsing phase. This makes the model focus more on the inherent features of the person rather than on extrinsic variations, thus reducing unnecessary computational burdens and improving the operational efficiency of the algorithm.

E. CC-ReID upon UAVs for Road Conditions Assessment and Monitoring Applications

Generally, the CC-ReID is applied to retrieve the special person with different clothes during a period, which can be used in smart cities for real-time road conditions assessment applications. Owing to the grayscale image reconstruction and

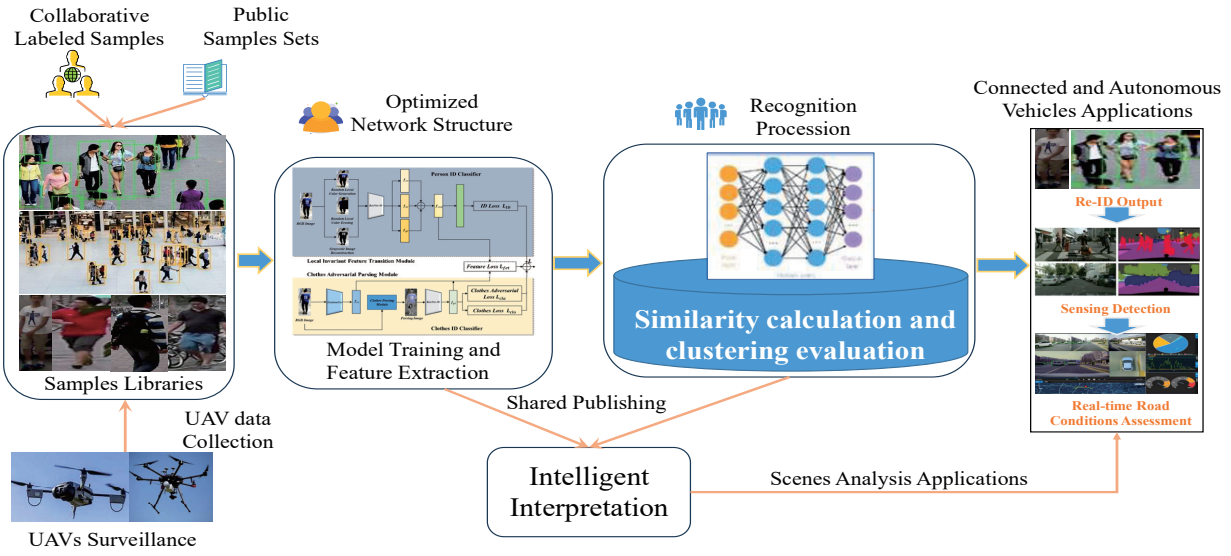

Fig. 4: The specific deployment for our proposed CC-ReID algorithm on the Intelligent Vehicles Control System.

parsing operations used in our method, the proposed CC-ReID algorithm has a high rate of retrieving special targets, achieving a time efficiency as low as $O(n)$. It can also be applied to smart terminals for road observations, such as unmanned aerial vehicles (UAVs) servers (Feisi X200) for security and risk assessment applications. The application procedure of the algorithm is illustrated in Fig. 4.

From Fig. 4, the main parts of the proposed algorithm deployment application are as follows: First, the collaborative labeled samples, public sample sets, and UAV-collected surveillance data comprised sample libraries that were used for model training and feature extraction. Subsequently, the optimized network structure is used for intelligent interpretation and the model recognition procedure based on UAVs, which is the similarity calculation and clustering evaluation. Finally, the deployed model is directly applied to connected and autonomous vehicle applications, including ReID output results, sensing detection, and real-time road conditions assessment.

Specifically, ReID tasks are evaluated by the CMC [30], which indicates the percentage of the real match in the list. For each probe, $p_i \in P$, all gallery images are ranked based on the similarity probabilities. In this case, once similarity probabilities are computed, the time complexity of the ranking algorithm can be as low as $O(n)$. As we know, this type of ranking algorithm is widely applied in many mobile surveillance devices [9], [47], which maintain limited computing ability. The Feisi X200 UAVs series mobile analysis server is an example of this kind of device, which is made by RflySim Platform (https://rflysim.com/).

In addition, to further validate the feasibility of deploying the model proposed in this work on UAVs servers, we have conducted the simulation experiments on the existing CC-ReID datasets (e.g., LTCC, PRCC, and CCVID) from the UAVs' viewpoints to verify the efficiency of instances matching. We have paid special attention to the efficiency of the model for instances matching, i.e., the number of instances matched in unit time. It indicates that the model can successfully match about 82371, 7300, and 241109 instances per minute on LTCC, PRCC, and CCVID datasets, respectively, as shown in Table IX. This result also indicates that the model can process data efficiently and is suitable for deployment on UAVs platforms, confirming our model's O(n) efficiency and deployment potential.

TABLE IX: Simulation results of instances matching efficiency from the perspective of UAVs.

Datasets	Every Minute		
	LTCC	PRCC	CCVID
Matched Instances	82371	7300	241109

V. Conclusions and Future Work

In this paper, we presented a data enhancement method called LIFTCAP, which is designed for the CC-ReID task. The proposed framework is equipped with LIFT and CAP models, which are used to extract invariant features and determine adversarial parsing contour differences between clothing styles. The results showed that our algorithm achieved competitive performance and could be deployed in an intelligent vehicle control system for real-time road conditions assessment applications.

Despite the novelty of our work and its superior performance, there is still room for improvement. For example, our algorithm for the fusion strategy of local invariant features must be further improved. In the future, we will consider image contour segmentation approaches for color transformation in real-time city road risk assessment.

References

[1] A. Aggarwal, "Enhancement of gps position accuracy using machine vision and deep learning techniques," *J. Comput. Sci.*, vol. 16, no. 5, pp. 651–659, 2020.

[2] V. Hassija, V. Chamola, V. Saxena, V. Chanana, P. Parashari, S. Mumtaz, and M. Guizani, "Present landscape of quantum computing," *IET Quantum Communication*, vol. 1, no. 2, pp. 42–48, 2020.

[3] X. Gu, H. Chang, and B. e. Ma, "Appearance-preserving 3d convolution for video-based person re-identification," in *European Conference on Computer Vision*. Springer, 2020, pp. 228–243.

[4] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Temporal complementary learning for video person re-identification," in *European Conference on Computer Vision*. Springer, 2020, pp. 388–405.

[5] J. Qian, M. Pan, and W. e. Tong, "Urrnet: A unified relational reasoning network for vehicle re-identification," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 9, pp. 11 156–11 168, 2023.

[6] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 2138–2147.

[7] T. Kumari, V. Guleria, P. Syal, and A. K. Aggarwal, "A feature cum intensity based ssim optimised hybrid image registration technique," in *2021 International Conference on Computing, Communication and Green Engineering (CCGE)*. IEEE, 2021, pp. 1–8.

[8] K. Arora and A. K. Aggarwal, "Approaches for image database retrieval based on color, texture, and shape features," in *Handbook of research on advanced concepts in real-time image and video processing*. IGI Global, 2018, pp. 28–50.

[9] H. Niu, Z. Lin, K. An, X. Liang, Y. Hu, D. Li, and G. Zheng, "Active ris-assisted secure transmission for cognitive satellite terrestrial networks," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 2, pp. 2609–2614, 2022.

[10] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," in *International conference on artificial intelligence*, vol. 33, no. 01. AAAI, 2019, pp. 8126–8133.

[11] X. Qian, W. Wang, L. Zhang, F. Zhu, Y. Fu, T. Xiang, Y.-G. Jiang, and X. Xue, "Long-term cloth-changing person re-identification," in *Asian Conference on Computer Vision*. Springer, 2020.

[12] Q. Yang, A. Wu, and W.-S. Zheng, "Person re-identification by contour sketch under moderate clothing change," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 6, pp. 2029–2046, 2019.

[13] X. Jin, T. He, K. Zheng, Z. Yin, X. Shen, Z. Huang, R. Feng, J. Huang, Z. Chen, and X.-S. Hua, "Cloth-changing person re-identification from a single image with gait prediction and regularization," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2022, pp. 14 278–14 287.

[14] D. Maini and A. K. Aggarwal, "Camera position estimation using 2d image dataset," *Int. J. Innov. Eng. Technol.*, vol. 10, pp. 199–203, 2018.

[15] S. Zhang, H. Gu, K. Chi, L. Huang, K. Yu, and S. Mumtaz, "Drl-based partial offloading for maximizing sum computation rate of wireless powered mobile edge computing network," *IEEE Transactions on Wireless Communications*, vol. 21, no. 12, pp. 10 934–10 948, 2022.

[16] X. Gu, H. Chang, B. Ma, S. Bai, S. Shan, and X. Chen, "Clothes-changing person re-identification with rgb modality only," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2022, pp. 1060–1069.

[17] Z. Yang, M. Lin, X. Zhong, Y. Wu, and Z. Wang, "Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2023, pp. 1472–1481.

[18] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *International conference on artificial intelligence*, vol. 34, no. 07. AAAI, 2020, pp. 13 001–13 008.

[19] Y. Gong and L. e. Huang, "Person re-identification method based on color attack and joint defence," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2022, pp. 4313–4322.

[20] X. Jia, X. Zhong, M. Ye, W. Liu, and W. Huang, "Complementary data augmentation for cloth-changing person re-identification," *IEEE Transactions on Image Processing*, vol. 31, pp. 4227–4239, 2022.

[21] K. Han, S. Gong, Y. Huang, L. Wang, and T. Tan, "Clothing-change feature augmentation for person re-identification," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2023, pp. 22 066–22 075.

[22] S. Zhang, Q. Zhang, Y. Yang, X. Wei, P. Wang, B. Jiao, and Y. Zhang, "Person re-identification in aerial imagery," *IEEE Transactions on Multimedia*, vol. 23, pp. 281–291, 2021.

[23] C. Zhang, X. Liu, J. Xu, T. Chen, G. Li, F. Jiang, and X. Li, "An edge based federated learning framework for person re-identification in uav delivery service," in *2021 IEEE International Conference on Web Services (ICWS)*, 2021, pp. 500–505.

[24] P. Wang, B. Jiao, L. Yang, and Y. e. Yang, "Vehicle re-identification in aerial imagery: Dataset and approach," in *International Conference on Computer Vision*. IEEE, 2019, pp. 460–469.

[25] C. Dong, J. Zhou, Q. An, F. Jiang, S. Chen, L. Pan, and X. Liu, "Optimizing performance in federated person re-identification through benchmark evaluation for blockchain-integrated smart uav delivery systems," *Drones*, vol. 7, no. 7, p. 413, 2023.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *International conference on computer vision and pattern recognition*. IEEE, 2016, pp. 770–778.

[27] T. He and Y. Hu, "Fashionnet: Personalized outfit recommendation with deep neural network," *arXiv preprint arXiv:1810.02443*, 2018.

[28] S. Lin, C.-T. Li, and A. C. Kot, "Multi-domain adversarial feature generalization for person re-identification," *IEEE Transactions on Image Processing*, vol. 30, pp. 1596–1607, 2020.

[29] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3260–3271, 2020.

[30] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *International workshop on performance evaluation for tracking and surveillance*, vol. 3, no. 5. IEEE, 2007, pp. 1–7.

[31] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *International conference on computer vision*. IEEE, 2015, pp. 1116–1124.

[32] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *International conference on computer vision and pattern recognition*. IEEE, 2018, pp. 2285–2294.

[33] J. Xue, Z. Meng, K. Katipally, H. Wang, and K. Van Zon, "Clothing change aware person identification," in *International Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2018, pp. 2112–2120.

[34] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *European conference on computer vision*. Springer, 2018, pp. 480–496.

[35] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *International conference on Multimedia*. ACM, 2018, pp. 274–282.

[36] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *International conference on computer vision*. IEEE, 2019, pp. 3702–3712.

[37] X. Qian, Y. Fu, T. Xiang, Y.-G. Jiang, and X. Xue, "Leader-based multi-scale attention deep architecture for person re-identification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 2, pp. 371–385, 2019.

[38] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *International conference on computer vision and pattern recognition workshops*. IEEE, 2019, pp. 0–0.

[39] P. Hong, T. Wu, A. Wu, X. Han, and W.-S. Zheng, "Fine-grained shape-appearance mutual learning for cloth-changing person re-identification," in *International conference on computer vision and pattern recognition*. IEEE, 2021, pp. 10 513–10 522.

[40] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *International conference on computer vision and pattern recognition*. IEEE, 2019, pp. 9317–9326.

[41] Y. Huang, Q. Wu, J. Xu, Y. Zhong, and Z. Zhang, "Clothing status awareness for long-term person re-identification," in *International Conference on Computer Vision*. IEEE, 2021, pp. 11 895–11 904.

[42] Z. Cui, J. Zhou, Y. Peng, S. Zhang, and Y. Wang, "Dcr-reid: Deep component reconstruction for cloth-changing person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[43] Q. Vadis, J. Carreira, and A. Zisserman, "Action recognition? a new model and the kinetics dataset," *Joao Carreira, Andrew Zisserman*.

[44] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *International conference on computer vision and pattern recognition*. IEEE, 2018, pp. 7794–7803.

[45] Z. Zhang, L. Tran, X. Yin, Y. Atoum, X. Liu, J. Wan, and N. Wang, "Gait recognition via disentangled representation learning," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 4710–4719.

[46] J. Gao and R. Nevatia, "Revisiting temporal modeling for video-based person reid," *arXiv preprint arXiv:1805.02104*, 2018.

[47] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 318–328, 2020.