Self-Aligned Spatial Feature Extraction Network for UAV Vehicle Reidentification

Aihuan Yao, Jiahao Qi, and Ping Zhong, Senior Member, IEEE

Abstract— Compared with existing vehicle reidentification (VeID) tasks conducted with datasets collected by fixed surveillance cameras, VeID for an unmanned aerial vehicle (UAV) is still under-explored and could be more challenging. Vehicles with the same color and type show extremely similar appearances from the UAV’s perspective so that mining fine-grained characteristics becomes necessary. Recent works tend to extract distinguishing information by regional features and component features. The former requires input images to be aligned and the latter entails detailed annotations, both of which are difficult to meet in UAV application. To extract efficient fine-grained features and avoid tedious annotating work, this letter develops an unsupervised self-aligned network consisting of three branches. The network introduced a self-alignment module to convert the input images with variable orientations to a uniform orientation, which is implemented under the constraint of a triple loss function designed with spatial features. On this basis, spatial features, obtained by vertical and horizontal segmentation methods, and global features are integrated to improve the representation ability in embedded space. Extensive experiments are conducted on UAV-VeID dataset, and our method achieves the best performance compared with recent reidentification (ReID) works.

Index Terms— Self-alignment module, spatial features, unmanned aerial vehicle (UAV) reidentification (ReID), vertical and horizontal segmentation.

I. INTRODUCTION

BEFORTING from the rapid development of microelectronics, navigation and communication technologies, unmanned aerial vehicle (UAV) technology has made great progress, which are widely used in both military and civil fields, including air reconnaissance, patrol security, and urban planning. The high-quality optical images obtained by UAV are used for computational analysis to realize object detection and scene classification [1], [2], [3], which is of great significance in the field of intelligent security. Vehicle reidentification (VeID), aiming at locating query vehicles from gallery set accurately, works as a key technology for social public security and smart city construction.

Manuscript received 24 December 2021; revised 23 June 2022 and 28 October 2022; accepted 6 January 2023. Date of publication 9 February 2023. This work was supported in part by the Natural Science Foundation of China under Grant 61971428.

(Author: Ping Zhong.)

Aihuan Yao was with the National Key Laboratory of Science and Technology on Automatic Target Recognition, National University of Defense Technology, Changsha 410073, China. She is now with Powerchina Zhongnan Engineering Corporation Ltd., Changsha 410014, China (e-mail: yaoah@nudt.edu.cn).

Jiahao Qi and Ping Zhong are with the National Key Laboratory of Science and Technology on Automatic Target Recognition, National University of Defense Technology, Changsha 410073, China (e-mail: qiijiahaor1996@163.com; zhongping@nudt.edu.cn).

Digital Object Identifier 10.1109/LGRS.2023.3237823

Fig. 1. Samples of UAV-VeID datasets. The first four images and the last four images come from different vehicles in each row.

Existing reidentification (ReID) works mainly focus on datasets collected by surveillance cameras, which possess fixed locations and limited viewpoints. It is the key to mine distinguishing characteristics for VeID. Many fine-grained feature extraction methods have emerged consequently, which can be generally divided into three categories: attention mechanism-based [4], regional feature-based [5], and component feature-based [6], [7], [8]. The attention mechanism can help models highlight valid features and has been widely used for image processing. Considering the imbalance of discriminative features in different spatial locations and different channels, spatial and channel attention network (SCAN) was proposed [4]. What’s more, as the local region conveys more distinctive visual cues, the region-aware deep model (RAM) also extracts features from a series of regions [5], to encourage the deep model to learn more discriminative features. Besides, some of the works utilize detailed annotation to find out key the positions to highlight effective component characteristics [6], including window, light, and brand.

VeID for UAV remote sensing images is still in the exploration stage in terms of methods and datasets. Teng et al. [9] contributed a novel dataset called UAV-VeID, as shown in Fig. 1, and proposed a viewpoint adversarial training strategy and a multiscale consensus loss to promote the robustness and discriminative power of learned deep features simultaneously. From Fig. 1, we can see that vehicle images sharing the same color and type demonstrate excessive similarity but variable vehicle orientation, which raises a high demand for the efficient feature extraction method. Whereas regional feature-based methods and component feature-based methods have the limitations of high dependence on input alignment and detailed annotations, respectively, which are not suitable for the ReID in UAV applications.

According to the above analysis, this letter proposes a self-aligned spatial feature extraction network (SANet), which can make a deep network focus on feature extraction among a
certain region and improve its representation ability. We introduce a self-alignment module to align input images, which makes it feasible to segment images according to a certain standard without viewpoint annotations. To achieve this purpose, we design a network with three branches and respond to metric loss functions. On the one hand, the network can integrate global and spatial features, for the sake of extracting fine-grained embedding features. On the other hand, loss functions designed for spatial features make the self-alignment module work.

II. METHODOLOGY

As mentioned, this letter aims to extract effective fine-grained features without relying on annotations. Therefore, we design the whole structure and the corresponding loss function to meet it.

A. Overall Framework

Our deep network architecture consists of three branches: global branch, upper/lower branch, and left/right branch. Obviously, the former is used to extract global features, and the latter two are used to extract spatial features in different directions. Input images have different deep features after three parallel network branches, and we concatenate them to obtain the final representation vector. The overall architecture is illustrated in Fig. 2.

1) Global Branch: It is worth mentioning that the first four blocks in Resnet-50 are selected as shared shallow networks. In addition, the global branch is composed of block-5 in Resnet-50, the pooling layer, and 1 × 1 convolution layer, which is used to reduce the dimension of embedded features.

2) Top/Down Branch: There is a self-alignment module for converting the input images to a uniform orientation between shared shallow networks and subsequent convolutional layers, a similar structure to block-5 in Resnet-50. The difference is that the step size is set to 1 for the sake of improving the resolution of feature maps. Then, we segment the output feature maps of the last convolutional layer evenly and horizontally to obtain the top and down spatial features.

3) Left/Right Branch: It possesses the same architecture as another spatial branch and shares the parameters of the self-alignment module with it. It is worth mentioning that the self-aligned module in both branches shares weights. In particular, we use vertical segmentation evenly to obtain the left and right spatial features.

B. Self-Alignment Module

Considering the gigantic diversity in vehicle orientations, it is fundamental to align input images for extracting spatial features. This section is introduced to implement spatial transform [10], which can convert the input to a uniform orientation without relying on viewpoint annotations. Consequently, it is feasible to divide vehicle areas in a fixed manner in the subsequent network.

The self-alignment module takes the output feature maps of shallow shared neural networks as input and outputs the self-aligned feature maps. It can be split into three parts: localization network, grid generator, and sampler. The calculation process could be described as follows. First, a customized localization network utilizes the input feature maps $U$ to generate the adaptive transform parameters $\theta$: $\theta = \text{func}_{\text{loc}}(U; W_{\text{loc}})$, where $\text{func}_{\text{loc}}$ is the localization network function and $W_{\text{loc}}$ is the corresponding parameters. In this letter, the localization network consists of two convolutional layers and two fully connected layers as shown in Fig. 2. Then, the grid generator creates a regular sampling grid $G = \{G_i\}$ according to $\theta$, where $G_i = (x_i^1, y_i^1)$ denotes the pixels on the output feature maps $V$. It is worth mentioning that, for each coordinate point of output feature maps $V$, the grid generator can generate a
particular location in $U$ to perform subsequent calculations. Therefore, the spatial transform between feature maps can be formulated as

$$
\begin{bmatrix}
  x_i^t \\
  y_i^t 
\end{bmatrix} = T_\theta(G_i) = \begin{bmatrix}
  \theta_{11} & \theta_{12} & \theta_{13} \\
  \theta_{21} & \theta_{22} & \theta_{23}
\end{bmatrix} \begin{bmatrix}
  x_i^o \\
  y_i^o \\
  1
\end{bmatrix}
$$

(1)

where $(x_i^t, y_i^t)$ and $(x_i^o, y_i^o)$ are the coordinates in the input feature maps $U$ and output feature maps $V$, respectively. Finally, $V$ could be obtained by means of differential image sampling, such as bilinear interpolation when the calculated coordinates are not integers. The main demand of spatial transformation from UAV’s perspective is to achieve self-alignment by rotating vehicle targets, and therefore, affine transformation is selected in this letter to meet this requirement. It is obvious that annotations are not necessary during the aforementioned computational process. The aligned feature maps are more suitable for UAV VeID tasks when extracting spatial features.

C. Objective Function

First, we regard the ReID task as a multiclassification problem and carry out classification loss constraints on the features extracted from the three branches, respectively,

$$
L^g_{\text{ID}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{ic} \log(p_{ic})
$$

$$
= -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{ic} \log \left(e^{W_{g}^{t}f_i^{t}+b_{ic}} \sum_{k=1}^{C} e^{W_{g}^{t}f_i^{k}+b_{ik}} \right).
$$

(2)

It shows the classification loss for the global branch, where $N$ and $C$ represent the number of training samples in mini-batch and the total number of vehicle IDs, respectively. Besides, $y_{ic}$ is a symbolic function, which equals 1 when the real class of the sample $i$ is equal to $c$, and 0 otherwise, $p_{ic}$ denotes the probability that the observed sample $i$ belongs to category $c$. Similarly, the classification loss for other branches, $L^{tri}_{\text{ID}}$ and $L^{top}_{\text{ID}}$, can be calculated by concatenating two spatial features, respectively.

Second, it is impossible to achieve self-alignment of input images only by triplet loss constraint on global features, which can be proved by related ReID methods who employ both triplet loss and classification loss. Therefore, in addition to the metric loss $L^g_{\text{ID}}$ for global features, this letter also sets corresponding metric loss for top, down, left, and right spatial features, which can be formulated as

$$
L^t_{\text{ID}} = \max \left( \| f_i^t - f_i^p \|_2^2 - \| f_i^o - f_i^n \|_2^2 + m, 0 \right)
$$

$$
L^l_{\text{ID}} = \max \left( \| f_i^l - f_i^p \|_2^2 - \| f_i^o - f_i^n \|_2^2 + m, 0 \right).
$$

(3)

(4)

The other three losses $L^d_{\text{ID}}, L^r_{\text{ID}}, L^{top}_{\text{ID}}$, and $L^{tri}_{\text{ID}}$ have a similar form, which encourages the distance of positive pairs to be smaller than negative sample pairs by a threshold $m$. Among them, $L^{tri}_{\text{ID}}$ is designed for features concatenated by global and spatial features. Using triplet loss functions to make the segmented spatial features gather in the embedded space, the self-aligned module can generate the adaptive parameters according to images with variable orientations, so as to guarantee the consistency of spatial features. In other words, the triplet loss functions designed for spatial features replace the viewpoint annotations to some extent.

D. Implementation Details

As mentioned above, pretrained Resnet-50 is selected as backbone network. Input images are resized to 256 $\times$ 256, and two methods of random erasing and color jitter are adopted. The margin of the triplet loss is set to 0.3, and the batch size is set to 32 $\times$ 8, indicating that four images of eight vehicle IDs are randomly selected in each batch. Adam optimizer with an initial learning rate of 0.0005 is adopted, which is adjusted according to cosine annealing strategy dynamically. Since the convergence result of the self-aligned module in the network has a great impact on network performance, the learning rate of the spatial transform module of spatial feature extraction branch is set as 0.0005 $\times$ 0.05 to avoid the failure of convergence due to excessive initial learning rate. In the process of training, considering that Resnet-50 loads pretraining parameters, we only optimize the parameters of the spatial transform module, 1 $\times$ 1 dimensional reduction convolution layer, and the full connection layer during the first ten epochs. Experiments are implemented based on NVIDIA 2080Ti GPU with 12-GB RAM.

III. EXPERIMENTS

In this section, abundant experiments have been carried out on UAV vehicles datasets to testify the effectiveness of our method. First, we briefly introduce necessary information about UAV-VeID [9] datasets and primary evaluation metric employed in this letter. Second, we analyze carefully the influence of the spatial features by conducting empirical studies, and the results of spatial transformation by visualization as well. At last, comparison experiments are conducted on UAV-VeID datasets and compared with recent works.

A. Datasets and Evaluation Metrics

1) Datasets: We first review necessary information about UAV vehicle datasets used in this letter, which is collected by UAV-mounted cameras.

UAV-VeID is constructed from video sequences captured by drones in different locations, backgrounds, and lighting conditions, such as highway intersections, road crossings, parking lots, etc. The flight altitude of UAV is between 15 and 60 m, and the vertical angle of the camera is between 40° and 80°, resulting in the variable sizes and perspectives of vehicle targets. The UAV-VeID dataset contains 41,917 images of 4,601 vehicles, of which the corresponding images of 1,797, 596, and 2,208 vehicles are used as training set, verification set, and test set, respectively.

2) Evaluation Metrics: During the test, the Euclidean distance between the embedded features of query set and gallery set is first calculated and sorted in ascending order, and then the distance metric matrix is returned. Because there is only
one ground truth match for a given query in UAV-VeID test set, we use only cumulative matching characteristics (CMC)-\(k\) to evaluate method performances, showing the probability of correct matching in the top-\(k\) ranked retrieved results. CMC-\(k\) calculates the mean top-\(k\) accuracy of all queries, which can be calculated as

\[
\text{Acc}_k = \begin{cases} 
1, & \text{if top-}k \text{ gallery samples contain the query identity} \\
0, & \text{otherwise} 
\end{cases}
\]

(5)

Since it only considers the first match in the evaluation process, it is accurate for datasets where there is only one ground truth match for a given query, such as UAV-VeID datasets.

### B. Model Analysis

In this section, we analyze the number of spatial feature blocks and then visualize the results of self-alignment.

The empirical study on spatial feature blocks of our SANet is implemented on UAV-VeID test set and the results are exhibited in Table I, where \(M\) denotes the number of blocks obtained during segmentation in the spatial branch of SANet, dim represents the dimension of integrating global and spatial features. It can be seen that there is little difference in ReID performance under the three kinds of spatial feature blocks, which can be interpreted as the high-dimensional features may contain more redundant information. Considering that in the process of VeID, CMC-1 evaluation metric has more practical application value, and the lower feature dimension is conducive to realizing VeID between larger databases. Therefore, we select SANet when \(M\) is 2 as the final network structure.

To testify whether the introduction of the self-alignment module achieved the expected effect, we randomly select several vehicle images of UAV-VeID test set and record the spatial transform parameters regressed by localization network. Then, the parameters are directly applied to the input image to verify the result of self-alignment, as illustrated in Fig. 3. The first two rows come from different vehicle IDs, and the last two rows demonstrate that the self-alignment module also has a good effect for images with changeful views, and uniformly converts the input images to the “upper right” orientation.

### C. Comparison With Recent Works

As mentioned above, this letter uses CMC-\(k\) index to evaluate the performance of ReID algorithms on UAV-VeID test set and compares it with several typical algorithms, as shown in Table II. Among them, the baseline model has the same network structure as SANet except for the self-aligned module, so as to validate the effectiveness of the module. Obviously, our method (SANet) achieves the best CMC-1 accuracy of 74.94%, outperforming other compared methods. It is interesting to observe that, CN-Nets [11], designed for fine-grained image retrieval with a coarse-to-fine framework, show a better performance compared with RAM [5] and SCAN [4]. Among them, RAM obtains regional features through horizontal segmentation directly, which is not suitable for the situation where vehicle orientations cannot be determined from the perspective of UAV. SCAN takes advantage of attentional mechanism to extract regional features but performs poorly. This can be interpreted as algorithms designed for VeID tasks on traditional surveillance videos do not consider the viewpoint variety in UAV-VeID datasets. Viewpoint-scale consistency reinforcement (VSCR) [9] and AM+WTL [12] utilize perspective annotations and color prior information, respectively. The former uses adversarial learning to generate features that are robust to perspective changes, while the latter emphasizes the target and reduces the background through attention mask. However, both of them are not as effective as method proposed in this letter, which indicates that correct fine-grained feature extraction is more conducive
to distinguishing similar vehicles. In general, SANet relies on no detailed annotation and achieves the optimal performance of UAV VeID.

In addition to UAV-VeID, we also conduct experiments on vehicle re-identification for aerial image (VRAI) [8], another challenging ReID dataset from UAV perspectives. It has a larger number of images, and the test set contains 71,500 images with 6,720 IDs. The results show that our method performs 72.12% on mAP metric for the test set, which is lower than [8]. Considering that SANet does not rely on any attribute annotations except vehicle IDs, the experimental result is acceptable.

More intuitively, Fig. 4 shows the CMC curves of each algorithm on the UAV-VeID test set, which shows the probability of correct results in the top-k images in the ReID results. Therefore, the closer the CMC curve is to the upper left corner, the stronger the performance of the corresponding algorithm is. Fig. 4 shows that the effect of SANet algorithm is generally superior to other comparison algorithms. Specifically, the hit rate of SANet before rank-10 is higher than VSCR, and the performance of the two algorithms after rank-10 is similar.

At last, we demonstrate some VeID results on UAV-VeID in Fig. 5, where the baseline and the SANet are compared. As we can see, our proposed method has better capability to distinguish the match results from similar false positives in gallery set compared with our baseline model. It demonstrates that our model learns more efficient and robust vehicle features.

### IV. Conclusion

In this letter, we analyzed the characteristics of vehicle images from the perspective of UAV first. Aiming at the variability of vehicle target orientation, we introduced the self-alignment module and designed triplet loss with spatial features to realize the alignment of input images without annotations. On this basis, we formulated the vertical and horizontal segmentation methods to extract spatial features, and integrate global features and spatial features simultaneously to improve the ability of representation in the embedded space. Extensive experiments verified that our method achieved the best performance on UAV-VeID datasets compared with recent works. As for VRAI datasets, SANet failed to achieve the best performance, but on the other hand, our method has low dependence on dataset annotations. In general, our model proposed in this letter does not rely on any additional annotation but achieves great performance in terms of evaluation metrics, which is more potential to be applied to real-world scenarios.

### References

[1] G. Cheng, Y. Si, H. Hong, X. Yao, and L. Guo, “Cross-scale feature fusion for object detection in optical remote sensing images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 431–435, Mar. 2020.

[2] S. Zhang, G. He, H.-B. Chen, N. Jing, and Q. Wang, “Scale adaptive proposal network for object detection in remote sensing images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 864–868, Oct. 2019.

[3] A. E. Almeida and R. Da Silva Torres, “Remote sensing image classification using genetic-programming-based time series similarity functions,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1499–1503, Sep. 2017.

[4] S. Teng, X. Liu, S. Zhang, and Q. Huang, “SCAN: Spatial and channel attention network for vehicle re-identification,” in *Proc. Pacific Rim Conf. Multimedia*, 2018, pp. 350–361.

[5] X. Liu, S. Zhang, Q. Huang, and W. Gao, “Ram: A region-aware deep model for vehicle re-identification,” in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[6] B. He, J. Li, Y. Zhao, and Y. Tian, “Part-regularized near-duplicate vehicle re-identification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3997–4005.

[7] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J.-C. Chen, and R. Chellappa, “A dual-path model with adaptive attention for vehicle re-identification,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6132–6141.

[8] P. Wang et al., “Vehicle re-identification in aerial imagery: Dataset and approach,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 460–469.

[9] S. Teng, S. Zhang, Q. Huang, and N. Sebe, “Viewpoint and scale consistency reinforcement for UAV vehicle re-identification,” *Int. J. Comput. Vis.*, vol. 129, no. 2385, pp. 719–735, 2020.

[10] M. Jaderberg, K. Simonyan, and A. Zisserman, “Spatial transformer networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, Jan. 2015, pp. 2017–2025.

[11] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian, “One-shot fine-grained instance retrieval,” in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 342–350.

[12] A. Yao, M. Huang, J. Qi, and P. Zhong, “Attention mask-based network with simple color annotation for UAV vehicle re-identification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.