remote sensing

MDPI

Article

MRG-T: Mask-Relation-Guided Transformer for Remote Vision-Based Pedestrian Attribute Recognition in Aerial Imagery

Shun Zhang *, Yupeng Li, Xiao Wu, Zunheng Chu and Lingfei Li

School of Electronic and Information, Northwestern Polytechnical University, Xi'an 710129, China; liyupeng1006@mail.nwpu.edu.cn (Y.L.); wuxiao@mail.nwpu.edu.cn (X.W.); chuzunheng@mail.nwpu.edu.cn (Z.C.); llf7103@mail.nwpu.edu.cn (L.L.)
* Correspondence: szhang@nwpu.edu.cn

Abstract: Nowadays, with the rapid development of consumer Unmanned Aerial Vehicles (UAVs), utilizing UAV platforms for visual surveillance has become very attractive, and a key part of this is remote vision-based pedestrian attribute recognition. Pedestrian Attribute Recognition (PAR) is dedicated to predicting multiple attribute labels of a single pedestrian image extracted from surveillance videos and aerial imagery, which presents significant challenges in the computer vision community due to factors such as poor imaging quality and substantial pose variations. Despite recent studies demonstrating impressive advancements in utilizing complicated architectures and exploring relations, most of them may fail to fully and systematically consider the inter-region, inter-attribute, and region-attribute mapping relations simultaneously and be stuck in the dilemma of information redundancy, leading to the degradation of recognition accuracy. To address the issues, we construct a novel Mask-Relation-Guided Transformer (MRG-T) framework that consists of three relation modeling modules to fully exploit spatial and semantic relations in the model learning process. Specifically, we first propose a Masked Region Relation Module (MRRM) to focus on precise spatial attention regions to extract more robust features with masked random patch training. To explore the semantic association of attributes, we further present a Masked Attribute Relation Module (MARM) to extract intrinsic and semantic inter-attribute relations with an attribute label masking strategy. Based on the cross-attention mechanism, we finally design a Region and Attribute Mapping Module (RAMM) to learn the cross-modal alignment between spatial regions and semantic attributes. We conduct comprehensive experiments on three public benchmarks such as PETA, PA-100K, and RAPv1, and conduct inference on a large-scale airborne person dataset named PRAI-1581. The extensive experimental results demonstrate the superior performance of our method compared to state-of-the-art approaches and validate the effectiveness of mask-relation-guided modeling in the remote vision-based PAR task.

Keywords: pedestrian attribute recognition; aerial imagery; relation modeling; masked attention mechanism

Citation: Zhang, S.; Li, Y.; Wu, X.; Chu, Z.; Li, L. MRG-T: Mask-Relation-Guided Transformer for Remote Vision-Based Pedestrian Attribute Recognition in Aerial Imagery. *Remote Sens.* **2024**, *16*, 1216. https://doi.org/10.3390/rs16071216

Academic Editor: Massimiliano Pepe

Received: 4 January 2024
Revised: 23 March 2024
Accepted: 26 March 2024
Published: 29 March 2024

1. Introduction

The Pedestrian Attribute Recognition (PAR) task [1] is dedicated to predicting multiple pedestrian attributes, such as gender, age, and body shape, as semantic descriptions for a single pedestrian image extracted from surveillance videos and aerial imagery. Nowadays, with the rapid development of consumer Unmanned Aerial Vehicles (UAVs), the use of UAV platforms for visual surveillance has become a necessary supplementary strategy to traditional fixed-camera position surveillance. UAV-based remote visual surveillance has received widespread attention from industry and academia due to its significant advantages such as flexibility and mobility, emergency response capabilities, cost effectiveness, remote operation, and automation. Different from the normal images captured by the traditional fixed cameras, aerial images have the following characteristics: (1) Viewing

angle and scale: UAV aerial images usually have a higher viewing angle and a wider scale range, while normal images may have a closer-to-the-ground viewing angle and a more localized scale. (2) Background environment: Aerial images may capture a broader background environment, such as city streets and suburban areas, while normal images may focus on specific scenes such as indoors or streets. (3) Image quality and clarity: Aerial images have higher image resolution and clarity because UAVs usually carry advanced camera equipment. Since the high-level semantic attributes are more robust to changes in viewpoint and diverse viewing conditions, remote vision-based pedestrian attributes in aerial imagery can be widely applied in person re-identification [2,3], pedestrian detection and tracking [4,5], and person retrieval [6,7].

Although it may seem straightforward to state, recognizing pedestrian attributes in real-world surveillance scenarios is still an extremely challenging task owing to three relation factors: (1) *Inter-region relations*: The spatial data redundancy and cross-region correlation in pedestrian images make it significantly difficult to extract distinct patterns and learn reliable image representations under background distraction. (2) *Inter-attribute relations*: A pedestrian image usually contains multiple attributes and some attributes are closely related to each other. (3) *Cross-modal Region-attribute mapping relations*: Pedestrian attributes may correspond to different parts of the image based on their semantic characteristics and some certain attributes may only be relevant to local regions. All these complex relations pose significant challenges in training an effective attribute recognition model.

To alleviate the above-mentioned issues, it is desirable to fully explore the inter-region, inter-attribute, and region-attribute mapping relations in the PAR task. For the *inter-region relation modeling* as shown in Figure 1a, the regions of the human body are related to each other; for example, short sleeves and shorts are mutually pushed from the background of the image. We need to eliminate the spatial data redundancy under background distraction and analyze long-range dependencies between regions. For example, in the process of identifying specific attributes (like age and gender), our attention may concentrate on multiple local regions, including those around the head, arm, leg, upper body, and lower body, and associate long-distance relations among these regions. Since Convolutional Neural Networks (CNNs) are struggling with learning long-distance dependencies, some existing methods adopt atrous convolutions [8], Feature Pyramid Networks (FPN) [9], Long Short-Term Memory (LSTM) [10,11], and Graph Convolutional Networks (GCN) [12–14] to alleviate the aforementioned limitation. Recently, some recent works [15] apply Vision Transformer (ViT) [16] as a feature extractor due to its capacity to capture long-distance relations between regions. In contrast, we construct a Masked Region Relation Module (MRRM) to focus on precise spatial attention regions to extract more robust features with masked random patch training.

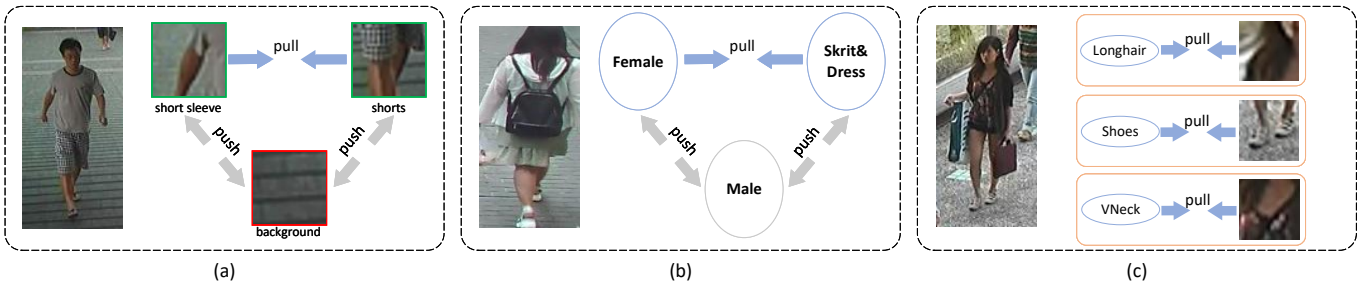

Figure 1. Pedestrian attribute recognition and relation modeling. (**a**) Inter-region relation modeling; (**b**) Inter-attribute relation modeling; and (**c**) Region-attribute mapping relation modeling.

For the *inter-attribute relation modeling*, as shown in Figure 1b, the ideal method will pull closer the relation between attributes "Female" and "Skirt&Dress" while simultaneously pushing farther the relation between attributes "Male" and "Skirt&Dress". It is clear that some specific attributes exhibit a strong semantic correlation. For example, the "Skirt" and "Dress" attributes are more likely to be associated with the attribute "Female"

than "Male". Inspired by this phenomenon, many methods have introduced this idea to learn the correlations among semantic attributes to improve pedestrian attribute recognition performance. The current mainstream methods use structured inference models, such as LSTM [17], GCN [13,14,18], and Vector-neuron capsule [19], to exploit the latent inter-attribute relation modeling. However, these methods do not consider the following two aspects: (1) There is a strong correlation between the existing attributes shown in the image, but the relationship between existing and non-existing attributes is weaker; (2) During the recognition process, many attributes are interfered with by attributes of the same category, resulting in inaccurate identification. In this paper, we construct a Masked Attribute Relation Module (MARM) which introduces self-attention to model attribute relations and explores its potential in semantic relation modeling. Meanwhile, MARM leverages masked attention to obtain more robust attribute relationships so that the interference of similar attributes can be alleviated and more accurate results and robust attribute relationships can be obtained.

For the *cross-modal region-attribute mapping relation modeling*, as shown in Figure 1c, there are mapping relations between regions and attributes. And some attributes may only be located in a small part of regions. For example, when we observe whether a person has "Glasses" or "Hats", we directly focus on the head region to make the mapping alignment between regions and attributes. Some existing methods attempt to address the problem with human pose estimations [20], leverage region proposal results [21,22] to identify related regions, or employ weekly supervised localization with an attention mechanism [23,24]. In this paper, we design a Region and Attribute Mapping Module (RAMM) based on the improved self-attention and cross-attention mechanism for modeling the relations between spatial features and semantic attributes and achieving more accurate mapping through mutual learning.

Although some existing methods mentioned some of the above three relations in their papers implicitly, it is desirable to jointly explore the inter-region, inter-attribute, and region-attribute mapping relations in one unified framework through a more competitive algorithm. In this paper, we explicitly exploit the inter-region, inter-attribute, and region-attribute mapping relations simultaneously with a novel Mask-Relation-Guided Transformer (MRG-T) framework that consists of three relation modules to fully exploit spatial and semantic relations in the model learning process. We construct a Masked Region Relation Module (MRRM) to extract more robust features using the ability of Transformer encoder layers and the masked random patch training strategy to establish correlations among regions through global attention to all regions. To delve into the connections between attributes, we present a Masked Attribute Relation Module (MARM) accompanied by an attribute label masking technique enabling the semantic capture of attribute relations. To learn the cross-modal alignment between spatial regions and semantic attributes, we finally design a Region and Attribute Mapping Module (RAMM) based on the self-attention and cross-attention mechanism. We analyze the contribution of each component in the proposed algorithm and demonstrate the effectiveness of relation modeling on three popular datasets. The experimental results show competitive performance on pedestrian attribute recognition.

Our work is distinctive from existing relation-based PAR works that consider only one relation factor, such as GRL [10], RC and RA [11] for inter-region relations; IAA-Caps [19] for inter-attribute relations; and LGNet [21], ALM [22]; and VTB [24] for cross-modal region-attribute mapping relations. Moreover, unlike some other approaches [13,14,18] that attempt to fuse these relation factors implicitly, our work explicitly and systematically presents to model and integrate the three types of relationships into one unified framework with the proposed mask-relation-guided Transformer. SCRL [25] is the most similar work which also considers intra-attentions in images and attributes and inter-attentions of spatial–semantic relations. However, it utilized RNN and the traditional attention units and considered borrowing person identity information to improve PAR performance.

Consequently, we summarize the main contributions of this paper as follows:

- We propose a novel Mask-Relation-Guided Transformer (MRG-T) framework to mitigate the information redundancy dilemma and model the three inter-region, inter-attribute, and region-attribute mapping relations simultaneously in a unified framework for remote vision-based PAR.
- We construct three modules, MRRM, MARM, and RAMM, to fully explore spatial relations of regions, semantic relations of attributes, and mapping of regions and attributes, respectively. The modules take advantage of the Transformer encoder architecture for its ability to capture long-distance dependencies from the global view.
- We present masked random patch training and attribute label masking strategies for MRRM and MARM, respectively, to conduct long-range dependency modeling of inter-region relations and inter-attribute relations efficiently. The beneficial effect of mask attention in the relational modeling method is proven through experiments.
- Our method performs favorably against state-of-the-art methods on three PAR datasets (PETA, PA-100K, and RAP) using the same backbone architecture. Moreover, we conduct model inference on a large aerial person imagery dataset PRAI-1581. Ablation experiments and visualization results are presented to demonstrate the capability of the proposed method in mask-relation-guided modeling.

2. Related Works

In this section, we provide a brief overview of recent studies on pedestrian attribute recognition, Transformer model and mask-attention modeling, and pedestrian attribute recognition based on Transformer.

2.1. Pedestrian Attribute Recognition

In recent years, a growing interest has been devoted to studying intelligent aerial surveillance [26]. Although most researchers have achieved successful developments on the tasks of object detection [27], tracking [28], and person ReID [29,30], there is limited attention paid to the task of PAR, probably because there does not exist a large-scale publicly available PAR dataset facing real UAV surveillance scenarios [31]. Therefore, in this subsection, our focus is primarily on reviewing the relevant PAR works within the surveillance domain.

In earlier studies on pedestrian attribute recognition, attributes were predominantly modeled using manually designed features like color histograms and texture histograms [32,33]. Recently, pedestrian attribute recognition methods based on deep learning have experienced significant breakthroughs, and most works employ either CNN architectures or attention mechanisms to extract distinctive representations [34,35]. The work in [36] considered the PAR task as a multi-label classification problem and devised a weighted sigmoid cross-entropy loss to simultaneously recognize multiple attributes. HydraPlus-Net [34] was developed by incorporating multi-directional attention (MDA) modules, allowing for encoding multi-scale features from multiple levels.

With the progress of research, some works are increasingly concentrating on investigating the correlation between regions and attributes. The main techniques contain multi-task learning [37], recurrent learning [10,18], LSTM [11], part-based localization [21,22], graph convolutional network (GCN) [13,14] and so on. (1) *Multi-task learning.* Sarfraz et al. [37] presented a multi-task learning model to jointly predict the coarse view and learn specialized view-dependent multi-label attribute inference. (2) *RNN and LSTM.* Wang et al. [18] proposed an end-to-end encoder–decoder recurrent network, called the Joint Recurrent Learning (JRL), to jointly learn image-level context and attribute-level relations using an LSTM model. In [10], an end-to-end Grouping Recurrent Learning (GRL) model is presented to make use of the intra-group mutual exclusion and inter-group relation to enhance PAR performance. Zhao et al. [11] presented the end-to-end Recurrent Convolutional (RC) and Recurrent Attention (RA) models. The RC model explored correlations among different attribute groups with a convolutional LSTM unit, while RA mined both intra-group attention locality and inter-group attention correlations. Wu et al. [25] proposed a sequence

contextual relation learning (SCRL) method to capture relations using RNN and attention. Person identity information was also leveraged to improve PAR performance. (3) *Part-based localization.* Localization Guided Network (LGNet) [21] exploited the attribute-specific local features based on the spatial similarity between region proposals and attribute locations. The attribute localization module (ALM) in [22] performed attribute-specific localization to learn the regional features for each attribute at multiple scales in a weakly supervised manner. (4) *GCN.* The visual–semantic graph reasoning framework [13] exploited spatial relations between regions with a spatial graph and learned potential semantic relations between attributes with a semantic graph. The framework performed reasoning with the Graph Convolutional Network (GCN). The work in [14] proposed a Joint Learning of Attribute and Contextual relations (JLAC) model, which constructed an attribute graph learned by GCN and designed a contextual relation graph to explore the contextual relations among those regions. (5) *Vector-neuron capsule.* Wu et al. [19] proposed an inter-attribute aware network via a vector-neuron capsule, called IAA-Caps, for PAR to be aware of relations between attributes.

Although these previous methods proposed to exploit relation modeling, few of them systematically integrate and model all types of relations in one unified framework. To address this issue, we propose a Mask-Relation-Guided Transformer framework with the self-attention and mask-attention mechanism to explicitly model relations of inter-regions, relations of inter-attributes, and mapping relations of regions and attributes, respectively, to jointly explore spatial and semantic relations in an end-to-end framework.

2.2. Transformer Model and Mask-Attention Modeling

Transformer-based models [38] were originally widely used in the field of natural language processing (NLP) [39,40] and have recently demonstrated excellent performance on computer vision (CV) tasks [16,41,42]. Vision Transformers (ViT) [16] were presented to split the images into sequences of image patches and then a standard Transformer encoder was applied to handle the image classification problem. To tackle the object detection task, DETR [41] regarded object detection as a direct set prediction task and assigned direct set prediction via transformers and bipartite matching. For the object re-identification task, TransReID [42] proposed a pure transformer-based object ReID framework in which a jigsaw patch module (JPM) is designed to learn more robust features, and side information embeddings (SIEs) are introduced to incorporate non-visual clues. At the same time, in other CV tasks, such as image captioning [43], cross-view gait recognition [44], video object detection [45], and image fusion [46], Transformer has shown its outstanding potential. The achievements of Transformers are primarily due to their success in the self-supervision and self-attention mechanisms [47]. Self-supervision enables the training of intricate networks without the expense of manual labeling while also encoding valuable relationships between presented entities. On the other hand, self-attention considers the correlation of the input sequence (like patches or words) by learning the correlations between the tokens. Some approaches [48–51] showcase the Transformer architecture's capability in capturing relations within sequences. In this paper, we also leverage the Transformer encoder to explore relations of regions, relations of attributes, and mapping relations between regions and attributes, with its ability to capture long-distance dependencies from the global view.

Mask-attention modeling can be divided into masked language modeling and masked image encoding in the fields of NLP and CV, respectively. Masked language modeling methods, e.g., BERT [52] and GPT [53–55], have proven to be remarkably effective for pre-training in the field of NLP. In these approaches, a portion of the input sequence is masked and then the models are trained to predict the omitted content. Extensive evidence shows that these methods generalize well to a variety of downstream tasks. Motivated by the success of similar techniques in NLP, some recent approaches in computer vision have adopted Transformer-based methodologies [38]. ViT [16] leverages the masked patch prediction objective for preliminary self-supervision learning, while BEiT [56] proposes to predict discrete tokens. Most recently, MAE [57] has adopted a strategy of masking random

patches in the input image; it then proceeds to reconstruct the missing pixels. Motivated by these results, we present an MRRM to extract more robust and distinct patterns with masked random patch training and an MARM to predict unknown attributes with parts of attributes known during training. Tao et al. [58] presented a pixel-level supervision neural network (PSNet) and designed an attention-based feature separation module (AFSM) to guide the interaction and separation process of background information and smoke information. In this paper, we design an MRRM using masked attention based on a Transformer to separate background information and pedestrian information. Lin et al. [59] proposed a masked attention mechanism to pay more attention to the feature information of the global text by reordering the weights corresponding to positions. In contrast, the mask module MARM in our proposed MRG-T is used mainly to learn more robust relationships between attributes by masking attributes and predicting output during training.

2.3. Pedestrian Attribute Recognition Based on Transformer

Inspired by the success of the Transformer in NLP and CV, some existing methods attempt to apply the Transformer in the pedestrian attribute recognition task. For DRFormer [15], it was first attempted to adopt ViT as the feature extractor on the pedestrian attribute recognition task. The Transformer encoder is leveraged to extract vector embedding features from spatial and semantic information. In STDP [60], the self-attention is extracted by the Swin Transformer to learn the relationships between spatial regions, and a transformer decoder is added to understand the semantic relationships among the attributes. However, neither DRFormer nor STDP explicitly exploit the cross-modal correspondence modeling between semantic attributes and image regions. For VTB [24], an additional textual modality is introduced and the PAR task is formulated as a multi-modal problem. The image and text modalities are aimed at modeling cross-modal relations, but the uni-modal relation modeling in images and attributes is not extensively explored. For PARFormer [61], a transformer-based multi-task model is built to focus on the global perspective. A multi-attribute center loss is designed to aggregate different attributes to their respective centers, as well as a multi-view contrastive loss for exploiting the viewpoint information. However, PARFormer mainly takes the Transformer as the feature extractor and does not use its superior self-attention power for relation modeling to improve model robustness.

Different from these methods that explore the relation modeling implicitly and individually, we explicitly and systemically model inter-regions, inter-attributes, and region-attribute mapping relations with two uni-modal Transformer encoders and one cross-modal Transformer encoder. Moreover, we present the masked random patch training strategy and the attribute label masking strategy for MRRM and MARM to efficiently conduct long-range dependency modeling of inter-region relations and inter-attribute relations, respectively.

3. Proposed Method

In this section, we first describe the overall architecture of the Mask-Relation-Guided Transformer. After that, we discuss three relational modeling modules, including the Masked Region Relation Module, the Masked Attribute Relation Module, and the Region and Attribute Mapping Module. Finally, we present the details of the inference classifier and the employed loss function for MRG-T.

3.1. Overall Architecture

Figure 2 shows the architecture of our proposed Mask-Relation-Guided Transformer framework for pedestrian attribute recognition. Solid lines denote high correlation probability, and dashed lines mean low correlation probability in MARM. We introduce MRG-T in three main parts.

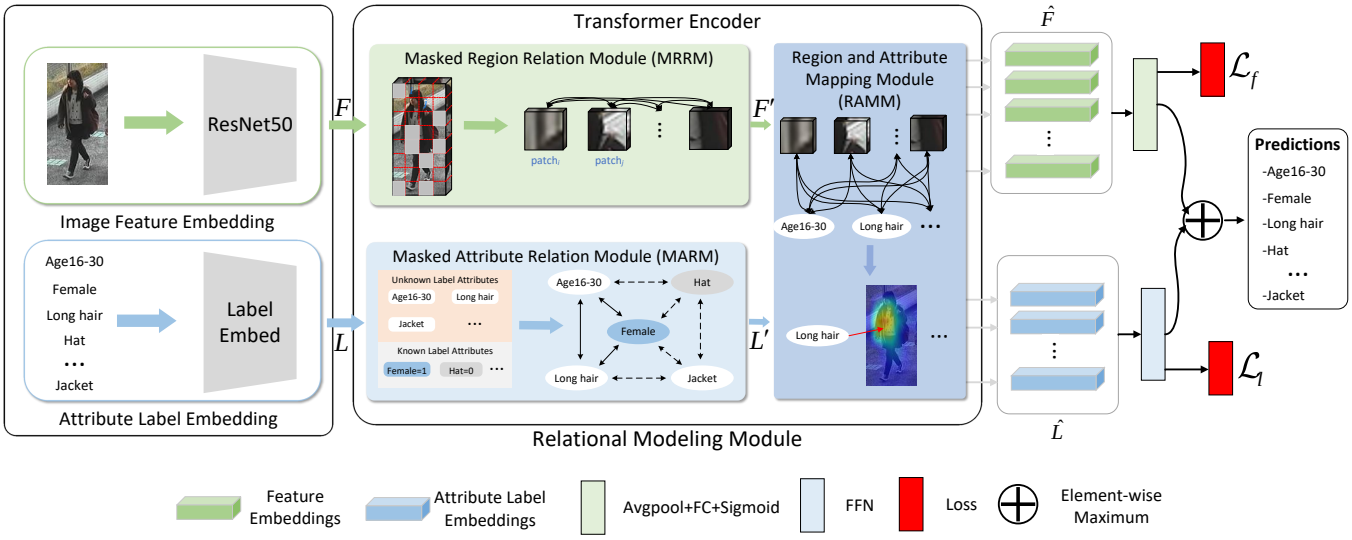

Figure 2. The overall architecture of our MRG-T is constructed with three modules: Masked Region Relation Module (MRRM), Masked Attribute Relation Module (MARM), and Region and Attribute Mapping Module (RAMM).

Feature and Label Embeddings. We leverage a CNN backbone (e.g., Resnet50 [62]) to extract visual feature embeddings F from input images. For attribute label embedding, we use an embedding layer to retrieve a set of label embeddings L from each attribute label.

Relational Modeling Modules. We mainly construct relational modeling with three modules: Masked Region Relation Module (MRRM) for inter-region relation modeling, Masked Attribute Relation Module (MARM) for inter-attribute relation modeling, and Region and Attribute Mapping Module (RAMM) for region-attribute mapping relation modeling. All three modules adopt the Transformer encoder architecture but differ in inputs and learning strategies. MRRM inputs the image feature embeddings F extracted by the backbone, a subset of features is randomly sampled, and the remaining features are masked with a masked random patch training strategy. MARM inputs the attribute label embeddings L and adopts the attribute label masking strategy to randomly eliminate a specific number of attribute labels. During training, it utilizes the ground truth of the remaining labels to predict the masked ones. MRRM and MARM output newly generated visual features F' and attribute label embeddings L', respectively. With F' and L' as inputs, RAMM learns the cross-modal alignment between spatial regions and semantic attributes and outputs the final visual features $\hat{F}$ and attribute label embeddings $\hat{L}$.

Inference Classifier and Loss Function. After feature and attribute label dependencies are modeled via the three relational modeling modules, we apply classifiers to make the final prediction. For visual feature embedding $\hat{F}$, we use the average pooling layer, the fully connected layer, and the activation function to obtain the final prediction $\mathcal{Y}_f$. For attribute label embedding $\hat{L}$, we use an independent feedforward network (FFN) for final prediction $\mathcal{Y}_l$. During training, we obtain the final loss with two binary cross-entropy loss functions, $\mathcal{L}_f$ and $\mathcal{L}_f$, by weighting the loss of the visual feature vector and the loss of the attribute label vector, respectively. During inference, we obtain the final prediction using the element-wise maximum.

3.2. Masked Region Relation Module (MRRM)

MRRM is built for inter-region relation modeling which aims to eliminate spatial data redundancy under background distraction and analyze long-range dependencies between regions so as to focus on precise spatial attention regions to extract more robust features.

The discriminative visual features fed into MRRM are obtained by extracting from each pedestrian image using a CNN backbone pretrained on ImageNet [63]. For a given pedestrian image, $x \in \mathbb{R}^{H\times W\times 3}$, the visual feature extraction can be formulated as

$$F = \mathcal{B}\{x; |\theta\}, \tag{1}$$

where $\mathcal{B}$ denotes the CNN backbone and θ encompasses all the trainable parameters associated with the CNN backbone. The output visual feature embeddings are $F \in \mathbb{R}^{h\times w\times d}$, in which h, w, and d represent the height, width, and channels, respectively.

We unfold the visual feature embeddings $F \in \mathbb{R}^{h\times w\times d}$ on each channel and then obtain $F = \{f_1, f_2, \ldots, f_N\}$, $f_i \in \mathbb{R}^d$ and $N = h \times w$. Subsequently, we proceed by randomly sampling a subset from N feature vectors and masking (i.e., removing) the remaining ones, following a uniform distribution. We denote the masking ratio as τ, which indicates the proportion of removed vectors. After masking random patches, we take masked tensor $\bar{F} = \{f_1, f_2, ..., f_n\}$ as regional feature embeddings to represent a local region mapped in the input image, where $n = (1 - \tau) \times N$. The random patch training strategy aims to effectively remove redundant information in the image space.

The self-attention mechanism of the Transformer encoder [16] (shown in Figure 3) is leveraged to capture the long-distance correlations between regional features, i.e., the weights of each regional feature relative to other regional features. With masked regional features $\bar{F}$ as inputs, we calculate the normalized attention coefficient α_{ij} between the i^{th} and j^{th} regional features, f_i and f_j, and then update each feature vector f_i to f_i' by computing the weighted sum and then passing through a nonlinear ReLU layer:

$$\alpha_{ij} = \text{softmax}\left(\left(W_r^Q f_i\right)^{\mathsf{T}}\left(W_r^K f_j\right)/\sqrt{d}\right), \tag{2}$$

$$f_i' = \text{ReLU}\left(\left(\sum_{j=1}^{n} \alpha_{ij} W_r^V f_j\right) + b_1\right) + b_2, \tag{3}$$

where W_r^Q, W_r^K, and W_r^V denote the query weight matrix, the key weight matrix, and the value weight matrix, respectively, and b_1 and b_2 represent bias vectors. We concatenate N_r Transformer encoder layers to capture long-range dependencies of iter-region relations, and each updated embedding is used as input for subsequent Transformer encoder layers. Note that the learned weight matrices $\{W_r^Q, W_r^K, W_r^V\} \in \mathbb{R}^{d\times d}$ are not shared between layers. With the implementation of multi-head attention, the modeling of the above-mentioned regional relationships can be represented as

$$F' = \text{MHA}_{\text{R}\to\text{R}}\left((f_i, \{f_1, ..., f_n\})|(W_r^Q, W_r^K, W_r^V)\right), \tag{4}$$

where $\text{MHA}_{\text{R}\to\text{R}}$ represents the calculated multi-head attentions between region features. The output feature embeddings of the Transformer encoder are denoted as $F' = \{f_1', f_2', \ldots, f_n'\}$, $f_i' \in \mathbb{R}^d$.

3.3. Masked Attribute Relation Module (MARM)

To address the issue of many attributes being interfered with by similar attributes, we construct the MARM using the self-attention mechanism and the attribute label masking strategy. The self-attention mechanism is introduced to model attribute relations and explores its potential in semantic relation modeling, and the attribute label masking strategy is used to obtain more robust attribute relationships to alleviate the interference of similar attributes.

Given one pedestrian image with μ annotated attribute labels (μ denotes the number of attributes), we extract the corresponding attribute label embeddings $L = \{l_1, l_2, \ldots, l_\mu\}$, $l_i \in \mathbb{R}^d$ and $i \in \{1, 2, \ldots, \mu\}$ with an embedded layer of dimensions $d \times \mu$. Considering the fact that the existing attributes shown in the image have strong correlations and the

relationship between existing and non-existing attributes is low, our proposed MARM incorporates such prior information by randomly masking a number of attribute labels and then utilizing the relations of the remaining attribute labels to predict the masked ones. As we randomly mask certain attribute labels, attribute labels may have three states ($s_i \in \{-1, 0, +1\}$): if labels are previously known to exist in the image, s_i is set to positive value $+1$; if labels are previously known not to exist, s_i is set to negative value -1; and if labels are unknown (or masked), $s_i = 0$. Similar to the work in [64], we incorporate the masked information by transferring it into state embeddings $m_i = \Phi(s_i)$, $m_i \in \mathbb{R}^d$. Transferred function $\Phi(\cdot)$ is implemented with a learned embedding layer of size $d \times 3$ in our MARM.

Given label embedding l_i and its corresponding state embedding m_i, we add the masked knowledge to the label embedding as follows:

$$\tilde{l}_i = l_i + m_i. \tag{5}$$

Thus, we obtain the final masked attribute label embeddings $\widetilde{L} = \{\tilde{l}_1, \tilde{l}_2, \ldots, \tilde{l}_\mu\}$, $\tilde{l}_i \in \mathbb{R}^d$. The masked attribute label embeddings $\tilde{L}$ are fed into the Transformer encoder to capture inter-attribute relations and learn the output L',

$$L' = \mathrm{MHA}_{\mathrm{A}\to\mathrm{A}}\Big((\tilde{l}_i, \{\tilde{l}_1, \ldots, \tilde{l}_\mu\})|(W_a^Q, W_a^K, W_a^V)\Big), \tag{6}$$

where $\mathrm{MHA}_{\mathrm{A}\to\mathrm{A}}$ represents the self-attention between attributes, and W_a^Q, W_a^K and W_a^V represent the learned weight matrices of query, key, and value about attributes, respectively.

In our experiments, the number of "unknown" labels for a pedestrian image is randomly set from 0.25μ to μ. By randomly masking various quantities of undisclosed labels during training, the model learns numerous potential label combinations and attribute relationships so as to predict the unknown labels.

3.4. The Region and Attribute Mapping Module (RAMM)

To learn the mapping of semantic attributes and spatial features, we propose the RAMM model with two separate Transformer encoders with the cross-attention mechanism (Shown in Figure 3). MRRM and MARM leverage the Transformer encoder based on the self-attention (SA) mechanism to capture correlations of sequences (like patches and words), while RAMM adopts the Transformer encoder based on the cross-attention (CA) mechanism to learn the cross-modal alignment between patches and attribute labels. Given the inputted feature embeddings F' and attribute embeddings L', we have

$$\hat{F} = \mathrm{MHA}_{\mathrm{R}\to\mathrm{A}}\Big((f_i', \{l_1', ..., l_\mu'\})|(W_f^Q, W_f^K, W_f^V)\Big), \tag{7}$$

$$\hat{L} = \mathrm{MHA}_{\mathrm{A}\to\mathrm{R}}\Big((l_i', \{f_1', ..., f_n'\})|(W_l^Q, W_l^K, W_l^V)\Big), \tag{8}$$

where $\mathrm{MHA}_{\mathrm{R}\to\mathrm{A}}$ and $\mathrm{MHA}_{\mathrm{A}\to\mathrm{R}}$ represent the cross-attention from regional features to attribute features and the cross-attention from attribute features to regional features, respectively. Wight matrices W_f^Q, W_f^K and W_f^V denote the parameters of $\mathrm{MHA}_{\mathrm{R}\to\mathrm{A}}$, while W_l^Q, W_l^K, W_l^V represent the parameters of $\mathrm{MHA}_{\mathrm{A}\to\mathrm{R}}$. Although the two Transformer encoders have the same architecture, the input query, keys, and values are different.

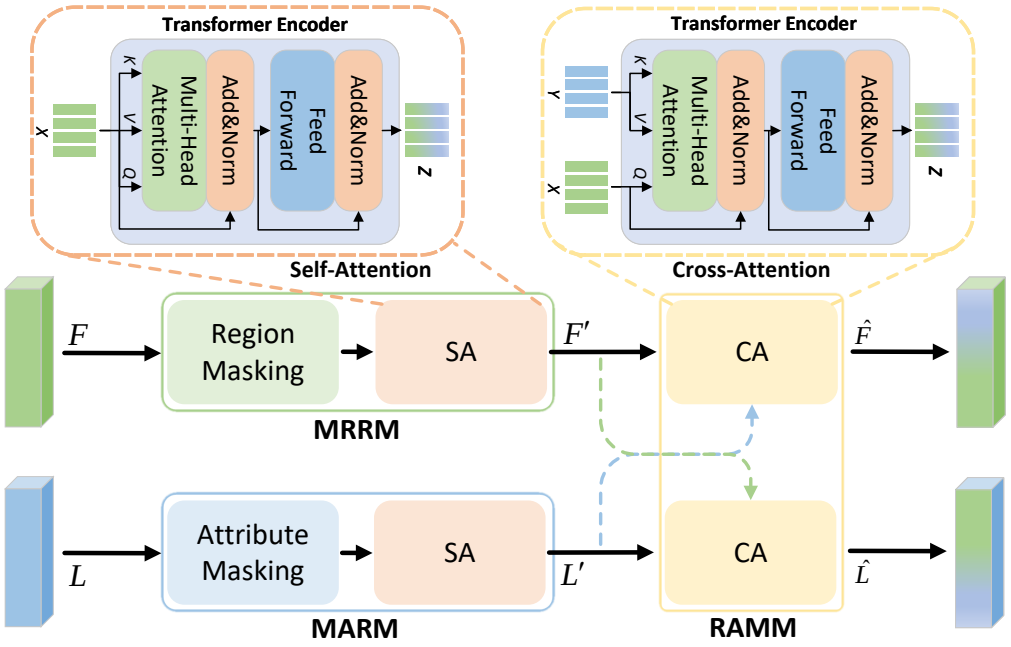

Figure 3. The proposed relational modeling modules are based on masking strategies and transformer encoders.

3.5. Final Classification and Loss Function

After the dependencies of regional features and attribute labels are modeled, we apply two classifiers to make the final predictions. We let $\hat{F} = \{\hat{f}_1, \hat{f}_2, ..., \hat{f}_n\}$ represent the output of the image feature embeddings and $\hat{L} = \{\hat{l}_1, \hat{l}_2, ..., \hat{l}_n\}$ be the output of the attribute embeddings. The image feature embeddings are transformed as $\hat{F} \in \mathbb{R}^{h \times w \times d}$ and then fed into an average pooling layer and a fully connected layer. The prediction result can be obtained after the final activation. For the attribute label embeddings, the prediction results are obtained through an independent feed-forward network (FFN) that includes a simple linear layer. We have the prediction results of both embeddings:

$$\mathcal{Y}_f = \sigma(\mathrm{FC}(\mathrm{avgpool}(\hat{F}))), \tag{9}$$

$$\mathcal{Y}_l = \mathrm{FFN}(\hat{L}) = \sigma((w \cdot \hat{L}) + b), \tag{10}$$

where $\mathrm{avgpool}(\cdot)$ represents the average pooling layer, $\mathrm{FC}(\cdot)$ represents the fully connected layer, $\sigma(\cdot)$ is a sigmoid function, w is a learned $1 \times d$ weight vector, and b is a bias vector.

Formally, each input image is coupled with its corresponding ground-truth attribute label $\widetilde{\mathcal{Y}} = \{\widetilde{\mathcal{Y}}^1, \widetilde{\mathcal{Y}}^2, \ldots, \widetilde{\mathcal{Y}}^M\}$, where M denotes the total number of attributes and $\widetilde{\mathcal{Y}}^m$ is a binary label indicating the presence of the m^{th} attribute if $\widetilde{\mathcal{Y}}^m = 1$, and $\widetilde{\mathcal{Y}}^m = 0$ otherwise. In our experiments, a weighted binary cross-entropy loss function is applied to train our model:

$$\begin{aligned}\mathcal{L}(\mathcal{Y}, \widetilde{\mathcal{Y}}) = -\frac{1}{M}\sum_{m=1}^{M} \gamma^m (\widetilde{\mathcal{Y}}^m \log(\sigma(\mathcal{Y}_i{}^m)) \\ + (1 - \widetilde{\mathcal{Y}}^m) \log(1 - \sigma(\mathcal{Y}_i{}^m))),\end{aligned} \tag{11}$$

where $\gamma^m = e^{-a_m}$ represents the loss weight for the m^{th} attribute to mitigate the imbalanced data issue in PAR datasets, a_m denotes the prior class distribution of the m^{th} attribute, and $\sigma(.)$ represents the sigmoid activation function.

For prediction results $\mathcal{Y}_f$ and $\mathcal{Y}_l$, we use two weighted binary cross-entropy losses, $\mathcal{L}_f = \mathcal{L}(\mathcal{Y}_f, \widetilde{\mathcal{Y}})$ and $\mathcal{L}_l = \mathcal{L}(\mathcal{Y}_l, \widetilde{\mathcal{Y}})$. Therefore, the total loss of training MRG-T is formulated as

$$\mathcal{L} = \lambda \mathcal{L}_f + \mathcal{L}_l, \tag{12}$$

where λ represents the trade-off weight parameter. During inference, the final prediction is obtained by aggregating the predicted results $\mathcal{Y}_f$ and $\mathcal{Y}_l$ with the element-wise maximum operation.

4. Experiments

In this section, we first introduce the datasets, evaluation metrics, and implementation details in our experiments. Then, we present the experimental results and conduct a data analysis regarding the contribution of each component in the proposed method.

4.1. Datasets and Evaluation Metrics

We construct experiments to validate our proposed method on three public PAR datasets (including PETA [33], PA100K [34] and RAPv1 [65]) and an aerial person imagery dataset PRAI-1581 [29].

PETA is a small-scale dataset for pedestrian attribute recognition, as shown in Figure 4a. It contains 8705 persons and consists of 19,000 images of individuals. Each image in the dataset is manually labeled with 65 attributes (61 binary and 4 multi-class). The dataset undergoes a random split into three subsets: 9500 images for the training set, 1900 images for the validation set, and 7600 images for the testing set. As per the evaluation in [33], 35 attributes that are chosen based on their positive ratios with a threshold set at 5% or higher are selected from the dataset. These chosen attributes are then employed for subsequent analysis and evaluation.

PA-100K is a recently introduced extensive dataset for pedestrian attribute recognition, as shown in Figure 4b. It consists of 100,000 pedestrian images gathered from 598 real outdoor surveillance cameras. These images have varying resolutions ranging from 50×100 to 758×454 pixels. As per the official setting [34], the dataset is partitioned into 80,000 images for the training set, 10,000 for the validation set, and the remaining 10,000 for the test set. Each image in the dataset is annotated with 26 commonly used attributes.

RAPv1 is a pedestrian attribute dataset captured from 26 indoor multi-camera surveillance scenarios, as shown in Figure 4c. It consists of 41,585 pedestrian images in total and their resolutions span from 36×92 to 344×554 pixels. Among these images, 33,268 images are selected for the training set, and 8317 are allocated for the testing set. Each image in the dataset is labeled with 72 fine-grained attributes, consisting of 69 binary attributes and 3 attributes with multiple classes. Similar to the work in [65], we also select 51 binary attributes that have a positive ratio exceeding 1% for fair comparison.

PRAI-1581 is a large-scale airborne person ReID dataset named Person ReID in Aerial Imagery, which consists of 39,461 images of 1581 person identities. As shown in Figure 4d, the images of the dataset are shot by two DJI consumer UAVs flying at an altitude ranging from 20 to 60 m above the ground, which covers most of the real UAV surveillance scenarios. It is worth noting that, due to variable altitude of flight, adjustable camera tilt angle, and free rotation of the fuselage, human images have different resolutions, perspectives, and postures in a single UAV. And since there are two independently controlled UAVs, the entire scene is more complex. Due to the difficulty of annotating attribute recognition datasets, it does not contain attribute labels, so we only explore inference and applications on it.

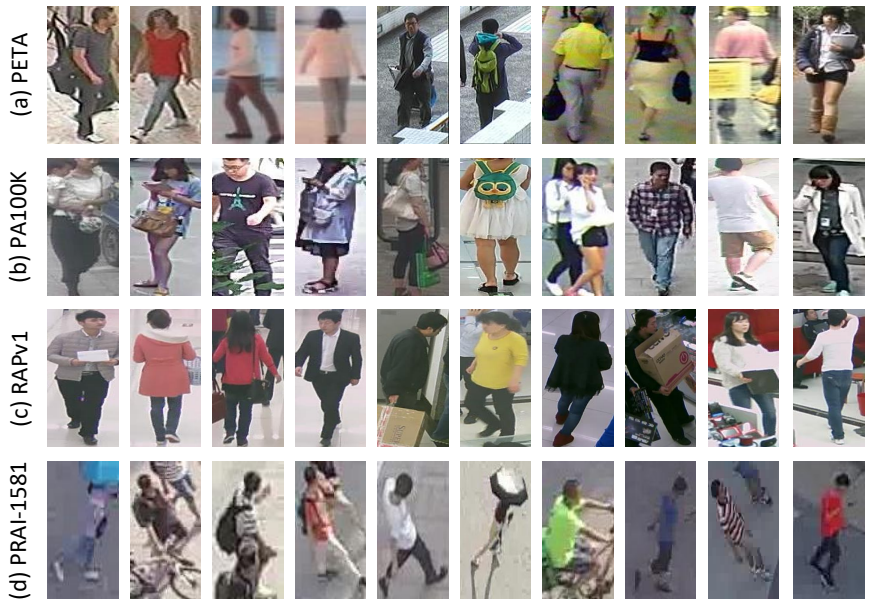

Figure 4. Some samples on various datasets, including three common datasets (i.e., PETA, PA100K, and RAPv1) and one airborne person dataset (i.e., PRAI-1581).

To assess the performance of the method, we employ five commonly used metrics in the PAR task to evaluate the pedestrian attribute recognition performance consisting of four instance-based evaluation metrics as well as a label-based evaluation metric. For the instance-based evaluation metrics, **Acc** (accuracy), **Prec** (precision), **Rec** (recall), and **F1** (F1-score) are used to evaluate the recognition performance of PAR methods at the instance level. In terms of the label-based evaluation metric, **mA** (mean accuracy) computes the classification accuracy for each attribute individually and then takes the average as the evaluation score. In contrast to the instance-based evaluation metrics, the label-based evaluation metric provides a more comprehensive assessment, taking into account all evaluation criteria, namely the average classification accuracy of both positive and negative samples. As a result, prior studies [14,19] have placed greater emphasis on **mA**, which offers a more robust and effective means of validating the proposed method.

4.2. Implementation Details

Experimental settings. The proposed MRG-T is implemented in the Pytorch framework [66] on two NVIDIA GTX 2080Ti GPUs and trained end-to-end. Each input image is resized to 224 $\times$ 224 for training. We employ the common data augmentation techniques including random horizontal mirroring, random rotation, and color jittering. The Adam optimizer [67] is leveraged as a stochastic gradient descent algorithm for training deep learning models, where $\beta_1 = 0.9$, $\beta_2 = 0.999$ and weight decay $\varepsilon = 0.0005$. The initial learning rate is set to 1e-4 and is reduced by a factor of 0.1 at the 20th, 30th, and 40th epochs. The total number of epochs for training is determined as 60, with the exception of the PA100K dataset, which requires a longer training period of 90 epochs due to its large-scale images [68]. For inference, all test images are also resized to 224 $\times$ 224 and then augmented only by normalization.

Backbone and Transformer encoder. ResNet50 [62] pre-trained on ImageNet [63] is leveraged as the CNN backbone to extract visual feature embeddings from input images. To improve the spatial resolution of the output feature maps, we omit the final down-sampling operation in the original ResNet50 architecture. Multiple attention heads are used to enhance the model's ability. To enhance the flexibility and capacity of the Transformer in capturing a wide range of patterns and dependencies in the input data, our model uses four attention heads [38]. We use $N_r = 3$ sequential connected Transformer encoder layers,

each of which applies the residual architecture with an embedding update and a layer norm. The effect of the number of these layers on model performance is discussed in ablation studies of Section 4.4.5.

4.3. Comparison with State-of-the-Art Methods

In Table 1, we present a performance comparison between our proposed MRG-T and existing state-of-the-art methods on the public datasets: PETA, PA-100K, and RAPv1. As depicted in Table 1, the compared methods are categorized into three groups:

- Holistic and localization methods, which localize the global body or local part of regions with the techniques of human pose estimation, region proposal extracting, or attention mechanisms, including DeepMAR [36], HPNet [34], PGDM [20], MsVAA [23], ALM [22], Baseline [69];
- Relation-based methods, which learn to model inter-region and inter-attribute relations by RNN, LSTM, GCN, or vector-neuron capsule, including JRL [18], RCRA [11], SCRL [25], SSC_{hard} [70], IAA-Caps [19];
- Transformer-based methods including VTB [24] and PARFormer [61].

Table 1. Performance comparison with state-of-the-art methods on the PETA, PA100K, and RAPv1 datasets. The first and second highest scores are represented by red font and cyan font, respectively. Best viewed in color.

Method	Backbone	PETA					PA100K					RAPv1				
		mA	Acc	Prec	Rec	F1	mA	Acc	Prec	Rec	F1	mA	Acc	Prec	Rec	F1
DeepMAR(ACPR15) [36]	CaffeNet	82.89	75.07	83.68	83.14	83.41	72.70	70.39	82.24	80.42	81.32	73.79	62.02	74.92	76.21	75.56
HPNet(ICCV17) [34]	InceptionNet	81.77	76.13	84.92	83.24	84.07	74.21	72.19	82.97	82.09	82.53	76.12	65.39	77.33	78.79	78.05
PGDM(ICME18) [20]	CaffeNet	82.97	78.08	86.86	84.68	85.76	74.95	73.08	84.36	82.24	83.29	74.31	64.57	78.86	75.90	77.35
MsVAA(ECCV18) [23]	ResNet50	84.35	78.69	87.27	85.51	86.09	80.10	76.98	86.26	85.62	85.50	79.75	65.74	77.69	78.99	77.93
ALM(ICCV19) [22]	ResNet50	85.50	78.37	83.76	89.13	86.04	79.26	78.64	87.33	86.73	86.64	81.16	67.35	74.97	85.36	79.39
Baseline(Arxiv21) [69]	ResNet50	84.42	78.13	86.88	85.08	85.97	80.38	78.58	87.09	87.01	87.05	80.32	67.28	79.04	79.89	79.46
JRL(ICCV17) [18]	AlexNet	85.67	–	86.03	85.34	85.42	-	-	-	-	-	77.81	-	78.11	78.98	78.58
RC(AAAI19) [11]	Inception_v3	85.78	–	85.42	88.02	86.70	-	-	-	-	-	78.47	-	82.67	76.65	79.54
RA(AAAI19) [11]	Inception_v3	86.11	–	84.69	88.51	86.56	-	-	-	-	-	81.16	-	79.45	79.23	79.34
SCRL(CSVT20) [25]	ResNet50	-	–	-	-	-	80.6	-	88.7	84.9	86.8	81.9	-	82.4	81.9	82.1
SSC_{hard}(ICCV21) [70]	ResNet50	85.92	78.53	86.31	86.23	85.96	81.02	78.42	86.39	87.55	86.55	82.14	68.16	77.87	82.88	79.87
IAA-Caps(PR22) [19]	OSNet	85.27	78.04	86.08	85.80	85.64	81.94	80.31	88.36	88.01	87.80	81.72	68.47	79.56	82.06	80.37
VTB(CSVT22) [24]	ResNet50	-	-	-	-	-	81.02	80.89	87.88	89.30	88.21	81.43	69.21	78.22	83.99	80.63
PARFormer(CSVT23) [61]	ResNet50	-	-	-	-	-	79.41	78.05	86.84	86.75	86.59	-	-	-	-	-
MRG-T w/o mask (Ours)	ResNet50	85.42	79.69	87.47	87.46	86.93	80.35	79.02	86.75	88.31	86.12	80.02	67.41	75.54	85.36	79.94
MRG-T(Ours)	ResNet50	86.22	79.86	86.53	89.51	87.09	81.24	79.92	87.91	89.61	86.66	82.10	69.16	77.67	86.48	80.41

MRG-T is our complete method, while MRG-T without the mask adopts our MRG-T but without mask attention. Recent advancements in PAR methods [24,61] have harnessed the capabilities of more powerful backbone architectures, including Vision Transformer (ViT-B) and Swin Transformer (Swin-B), often accompanied by self-attention or shift-window attention mechanisms. To ensure fair comparisons in this paper, we mainly list the experimental results of the methods with the ResNet50 backbone.

Comparison with holistic and localization methods. Overall, our method performs favorably against the state-of-the-art approaches. DeepMar makes the first attempt to incorporate CNN into the PAR task and it learns a multi-label classifier to simultaneously identify multiple attributes. However, the CNN model lacks the ability to capture global information and most of these methods have not considered it. On the contrary, MRG-T builds inter-region, inter-attribute, and region-attribute relations simultaneously. Moreover, compared to Baseline [69], the MRG-T achieves 1.80%, 0.86%, and 1.78% improvements in terms of the mA metric on the three public datasets.

Comparison with relation-based methods. It can be seen that our proposed MRG-T method achieves relatively better performance compared with other relation-based methods in terms of most metrics. Despite IAA-Caps obtaining higher performance in some criteria on PA100K, our method, MRG-T, still outperforms IAA-Caps on the other

datasets. For example, our proposed method outperforms IAA-Caps in mA by 0.95% and 0.38% on PETA and RAPv1, respectively. In addition, the IAA-Caps method is limited in the relatively higher GPU memory usage during the training phase [19].

Comparison with Transformer-based methods. In comparison to Transformer-based approaches, our method demonstrates commendable performance with the same backbone. Specifically, our method outperforms the recent Transformer-based approach, PARFormer [61], in terms of all metrics on the PA100K dataset. Compared to VTB on PA100K and RAPv1, our MRG-T achieves 0.22% and 0.67% improvement in mA, and 0.31% and 2.49% improvement in recall.

Comparison of applications on aerial imagery dataset. Since PRAI-1581 lacks attribute label-level annotations, we manually annotate the attributes of the sub-dataset and then perform model inference to evaluate the recognition performance. Figure 5 shows the recognition results of our method and the compared method ALM on several pedestrian images from PRAI-1581. For each example image, the correct and wrong predicted attributes are marked in green and red colors, respectively. The true, false, and missed numbers of predicted attributes are also given. The results show that our proposed MRG-T performs better than ALM, especially when some attributes are related. For example, the semantic association between "Shorts" and "ShortSleeve" in the first image and the association between "Longhair" and "Skirt" in the second image.

Figure 5. Qualitative evaluation of some pedestrian images from aerial imagery dataset PRAI-1581. The correct and wrong predictions are marked in green and red, respectively.

4.4. Ablation Studies

In this section, we conduct ablation studies to validate the effectiveness of each component proposed in our MRG-T method. These experiments are mainly conducted on the PETA, PA100K, and PRAI-1581 datasets.

4.4.1. Effectiveness of MRG-T

For quantitative analysis, we first compare the attribute-wise mA between ResNet50 (Backbone) and MRG-T (Ours) for all attributes on the PA100K dataset. Here, ResNet50 (Backbone) just employs the ResNet50 to extract visual feature embeddings and then trains a multi-label classifier to tackle the pedestrian attribute recognition issue. Figure 6 illustrates that MRG-T leads to a substantial performance improvement on almost all attributes. For some attributes, such as "Female", "Age", and "Upperstride", which require focusing on multiple regions of the image from the global view, our MRG-T achieves improvement on mA because the proposed relation modeling modules are able to capture long-range dependencies of multiple local regions. For some attributes related to specific spatial regions, such as "Hat" and "boots", our method undergoes more significant improvement because of our region-attribute mapping relation modeling with RAMM.

Figure 7 illustrates the qualitative attribute prediction results of two examples on the PETA dataset. The left column shows the raw input images, while the red, green, and blue tapes on the right illustrate the ground truth attribute labels, the attribute predictions by ResNet50, and the attribute recognition results by our proposed MRG-T, respectively. We can observe that MRG-T demonstrates a higher accuracy in identifying global attributes like "Age" and has a more accurate understanding of the relations between attributes, such as "Jeans" and "LeatherShoes", due to the exploration of correlations.

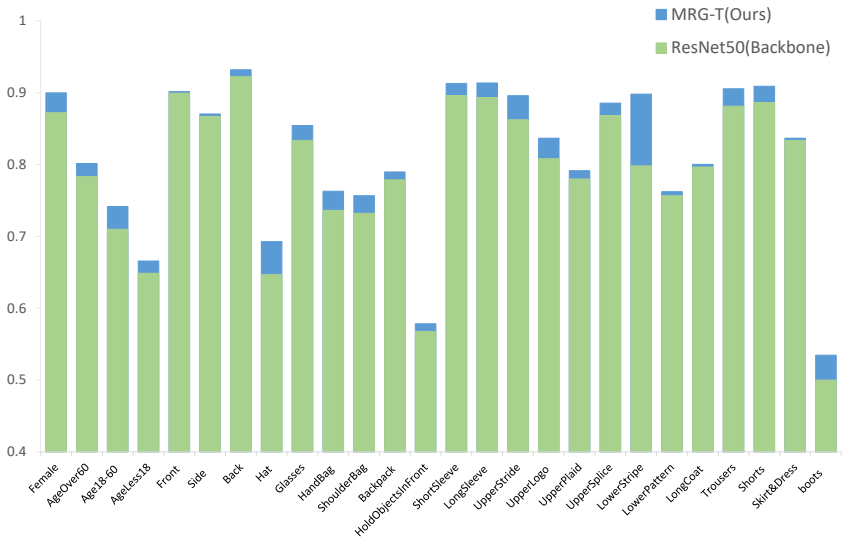

Figure 6. The mA comparison between ResNet50 (Backbone) and MRG-T (Ours) for all attributes on the PA100K dataset.

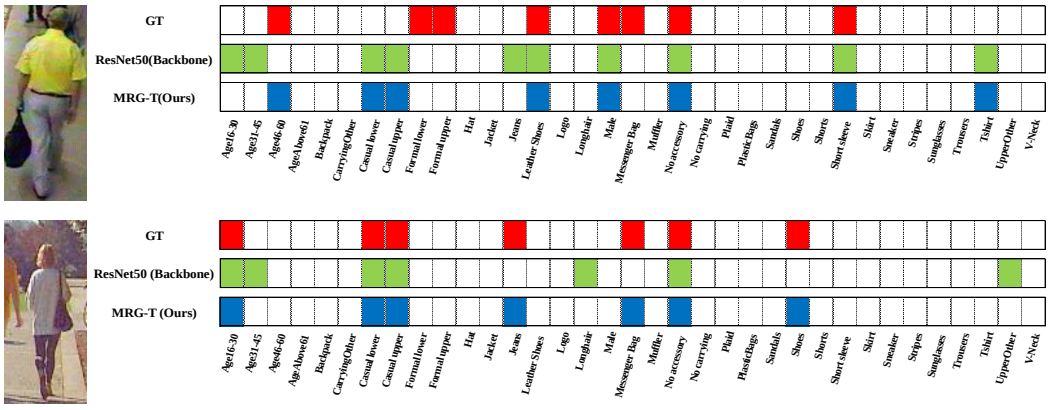

Figure 7. Attribute prediction results of ResNet50 (Backbone) and MRG-T (Ours) on the PETA dataset.

4.4.2. Benefit of Spatial Relations of Regions

In MRG-T, our MRRM uses a self-attention mechanism to acquire a wide range of inter-region relations by masking unnecessary and redundant information in the input image. With MRRM, the MRG-T model can improve by 0.39% in mA, by 0.41% in Acc, by 0.32% in Prec, by 0.79% in Rec, and by 0.45% in F1, as shown in Table 2. Meanwhile, we show the relations between regions learned by MRG-T through the cosine similarity of feature embeddings as shown in Figure 8. The numbers represent the remaining area of the image after masking, and the patch is 1–36 from left to right and top to bottom. It can be observed that the image regions representing unnecessary information (e.g., numbers 1, 5, 35, and 36) have no obvious relations with the body regions. Our MRRM mainly extracts the relations of regions representing the human body.

Table 2. Performance with or without the proposed MRRM in MRG-T on the PETA dataset. **Bold** text indicates the best performance.

Component	mA	Acc	Prec	Rec	F1
w/o MRRM	85.83	79.45	86.21	88.72	86.64
w MRRM	**86.22**	**79.86**	**86.53**	**89.51**	**87.09**

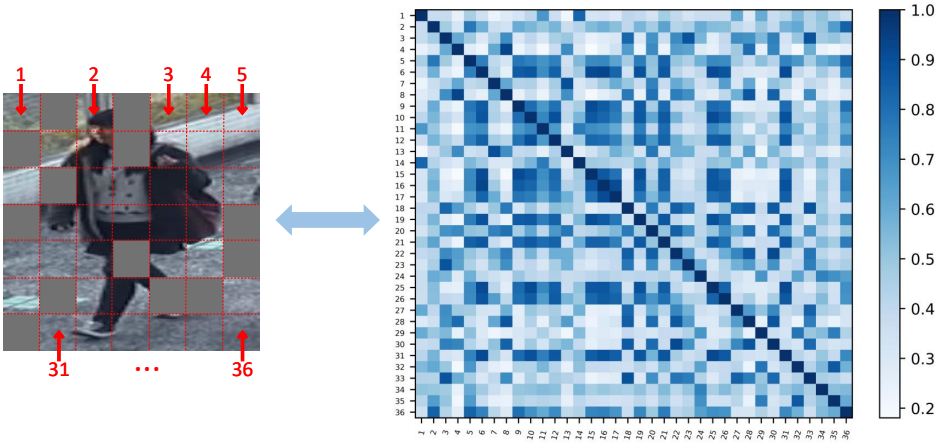

Figure 8. Illustration of the cosine similarity of feature embeddings before classifier layer in MRG-T. On the **left** is the masked image with patches counting from 1 to 36, and on the **right** is the visualization of inter-region relations. Best viewed in color.

4.4.3. Benefit of Semantic Relations of Attributes

In MRG-T, our MARM uses masked attention to obtain more robust attribute relationships. With MARM, the MRG-T model can improve by 0.63% in mA, by 0.25% in Acc, by 0.30% in Prec, by 1.19% in Rec, and by 0.18% in F1, as shown in Table 3. Table 4 shows the experimental results with different ratios of masking ratios in four simulation settings. When the masked ratio is 0.75, we achieve the best improvement in mA.

Table 3. Performance with or without the proposed MARM in MRG-T on the PETA dataset. **Bold** text indicates the best performance.

Component	mA	Acc	Prec	Rec	F1
w/o MARM	85.59	79.61	86.23	88.32	86.91
w MARM	**86.22**	**79.86**	**86.53**	**89.51**	**87.09**

Table 4. The mA of different masked attribute ratios in the MARM module. The **Bold** text indicates the best performance.

Masked Attribute Ratio	PETA			
	25%	**50%**	**75%**	**100%**
MRG-T (Ours)	85.88	85.90	**86.22**	85.68

Moreover, to explore the inter-attribute relationship, we compute the similarity between different attribute feature embeddings. Here, we employ cosine similarity ($Cos\langle\cdot,\cdot\rangle$) as a metric to gauge the inter-attribute relations. The cosine similarity value spans from -1 (indicating a strong negative correlation) to 1 (representing a strong positive correlation). Consequently, a larger cosine similarity value implies a stronger relationship between two distinct attributes. As illustrated in Figure 9, we calculate the cosine similarity of attribute embeddings in MRG-T on the test set of PETA and PA100K. The results demonstrate that the proposed MRG-T method effectively discerns the correlation among different attributes. Some attributes have higher relations, e.g., $Cos\langle LongCoat, boots\rangle$ is 0.9568, $Cos\langle Casuallower, Casualupper\rangle$ is 0.9710, and $Cos\langle Formallower, Formalupper\rangle$ is 0.9764. Additionally, for attributes with weaker correlations, such as Longhair and Male, MRG-T also reveals a negative correlation between the two attributes, e.g., $Cos\langle Longhair, Male\rangle$ is -0.6929. In Table 5, we provide the analysis results with respect to the impacts of MRRM and MARM. We select four common sets of attributes, including two sets of positive samples and two sets of negative samples. The results show that MRRM and MARM are

able to make the relationship between attributes more accurate. For example, the relation between Longhair and Skirt becomes stronger, and the relation between Casuallower and LeatherShoes becomes weaker.

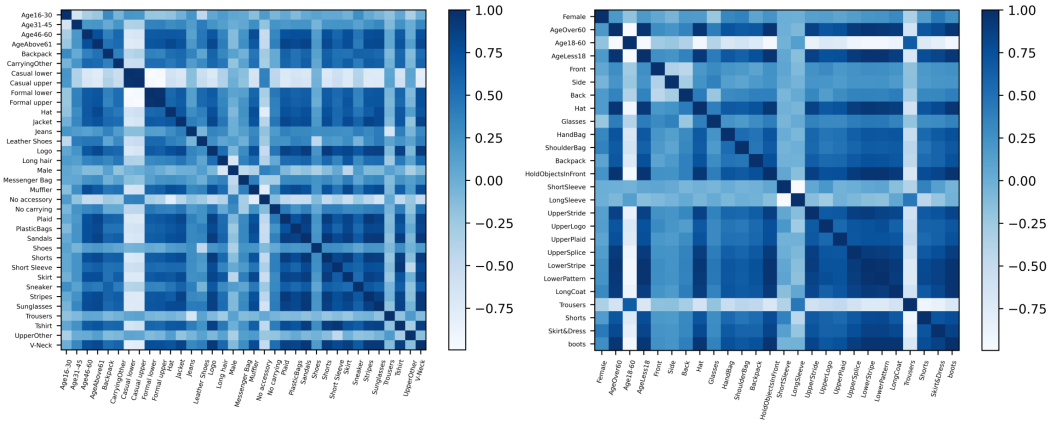

Figure 9. Illustration of cosine similarity of attribute embeddings in MRG-T on PETA and PA100K datasets. Data on the **left** are obtained from PETA, and on the **right** from PA100K. Best viewed in color.

Table 5. Relation analysis between attributes of MRG-T. Operation $Cos\langle\cdot,\cdot\rangle$ is short for cosine similarity, and $Cos\langle a, b\rangle \in [-1, 1]$.

Method		$Cos\langle Longhair, Male\rangle$	$Cos\langle Longhair, Skirt\rangle$	$Cos\langle Casuallower, LeatherShoes\rangle$	$Cos\langle Formallower, LeatherShoes\rangle$
MRRM	MARM				
-	-	−0.6578	0.6267	−0.5138	0.5299
-	✓	−0.6731	0.6476	−0.5598	0.5711
✓	✓	−0.6929	0.6998	−0.5739	0.5925

4.4.4. Mapping Visualizations of Regions and Attributes

In MRG-T, our RAMM models the relation mappings between spatial features and semantic attributes. We visualize the activation maps before the classifier layer to gain a deeper insight into how RAMM captures the mapping alignment between regions and attributes. We employ the widely used visualization technique, Grad-cam [71], for generating visualizations. Figure 10 shows that the proposed MRG-T can locate attributes at different regions according to its semantic characteristics. We illustrate more activation maps of attributes in Figure 11. When recognizing attributes of pedestrians, RAMM helps to locate the region that needs attention. For instance, when identifying the attribute "Hat", more attention is directed toward the head region, and when recognizing the attribute "Jeans", it pays close attention to the lower body.

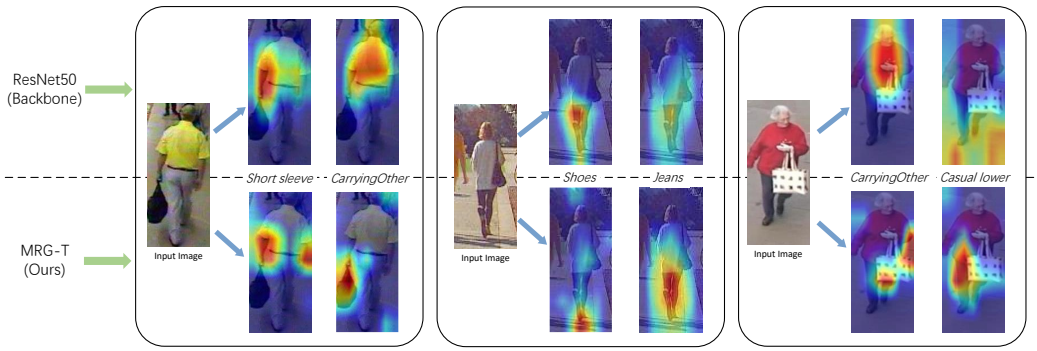

Figure 10. Visualization of the activation map of ResNet50 (Backbone) and MRG-T (Ours) models.

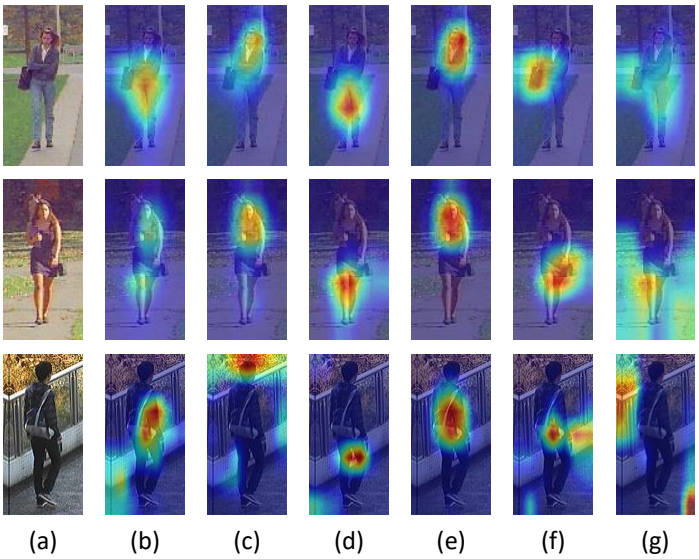

Figure 11. Activation map visualization of MRG-T on PETA dataset: (**a**) denotes raw image, and (**b**–**g**) represents activation map of attributes of **CarryingOther**, **Hat**, **Jeans**, **Longhair**, **Messenger Bag**, and **No accessory**, respectively.

4.4.5. Settings Analysis of MRG-T

We analyze the effect of transformer encoder and loss weights.

Effect on Transformer Encoder. The performance that MRG-T achieved can be affected by the layers of multi-head attention and the number of encoder layers in the Transformer, which play a crucial role in modeling correlations among embeddings of different regions and attributes. To assess the significance of multi-headed self-attention, we conduct experiments by changing the number of multi-head attention layers and observe that the mA improves as the multi-head attention layers increase. Due to resource constraints, we set the number of multi-head attentions to four in our experiments. Additionally, we explore the significance of the number of encoder layers, denoted as N_r. With the N_r set to two, three, and four, the mA values are 86.14%, 86.22%, and 86.12%, respectively. Consequently, we opt to set the number of encoder layers N_r to three in our experiments.

Effect of loss weights. In MRG-T, two loss functions are involved: $\mathcal{L}_f$ and $\mathcal{L}_l$, as shown in the Equation (12). Parameter λ represents the trade-off between $\mathcal{L}_f$ and $\mathcal{L}_l$. We evaluate the effect of weight parameter λ in Equation (12) as shown in Figure 12. When $\lambda = 0.0$, $\mathcal{L}_f$ is discarded and cannot exploit relations between different regions. With the injection of $\mathcal{L}_f$, mA improves significantly. MRG-T model attains its peak performance when λ is set to 0.2. Furthermore, if the value of λ is further increased, the performance experiences a gradual decline, which demonstrates the robustness of MRG-T with respect to λ.

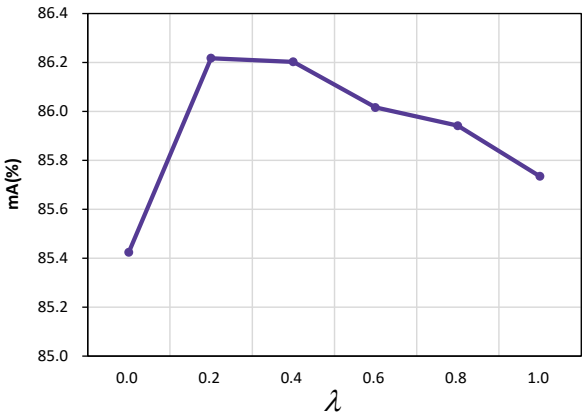

Figure 12. Effect of weight parameter λ in Equation (12) on the PETA dataset.

4.4.6. Ablation Studies on Aerial Image Dataset PRAI-1581

In this section, ablation studies are provided to systematically analyze the proposed MRG-T method on the PRAI-1581 dataset. Since PRAI-1581 is an aerial image person ReID dataset, it does not contain the attribute labels required by PAR. In this paper, we manually annotate 218 images, a total of 11 pedestrians, for ablation experiments. Four experimental settings shown in Table 6 are designed to analyze the influence of each module on retrieval effectiveness.

- T_0: Use only visual features extracted by ResNet50 and label embeddings retrieved by an embedding layer.
- T_1: Add visual mask-guided Transformer MRRM based on T_0 to model inter-region relations of images.
- T_2: Add label mask-guided Transformer MARM based on T_1 to model inter-attribute relations of labels.
- T_3: Add RAMM module to model region-attribute mapping relations.

Table 6 shows that our proposed MRG-T achieves remarkable performance in inference on aerial image datasets. Compared with T_0, after adding the Transformer of visual features MRRM to the T_1 model, mR increases by 9.31. After introducing T_2 to explore the relationship between attributes by MARM, the mR increases by 2.48. Finally, RAMM is introduced to model the mapping relationship between the image regions and label attributes, which improves by 1.35 compared to T_2 and improves by 13.14 compared with the initial T_0.

Table 6. Ablation studies of different compositions on the aerial image dataset PRAI-1581.

Ablation Model	Module			mA
	MRRM	MARM	RAMM	
T_0				80.13
T_1	✓			89.44
T_2	✓	✓		91.92
T_3	✓	✓	✓	93.27

4.5. Discussion

In this subsection, we provide more analysis and discussion on the superiority and the weaknesses of the proposed method compared to the existing literature.

We propose a novel framework named MRG-T to mitigate the image and attribute information redundancy dilemma by two different masking strategies and model the three inter-region, inter-attribute, and region-attribute mapping relations simultaneously in a unified framework for remote vision-based PAR. Compared with holistic and localization methods such as MsVAA [23], ALM [22], Baseline [69], etc., our proposed method first uses CNN to extract local features and then uses Transformer to model relationships while also obtaining a global receptive field. It displays better performance in the recognition of global attributes such as gender and age. Compared with relation-based methods, most current methods focus on inter-region or inter-attribute relations such as JRL [18], RCRA [11], SCRL [25], SSC_{hard} [70], and IAA-Caps [19]. Our proposed method models three relationships simultaneously, which is more comprehensive and specific. Compared with Transformer-based methods, such as VTB [24] and PARFormer [61], the model we propose uses Transformer to model explicit relationships, not just to obtain global features.

Although the method proposed in this paper achieves excellent performance, it has a relatively large number of parameters compared with some existing methods, such as ALM [22], RC [11], and RA [11]. Of course, the number of parameters has not increased significantly compared to VTB [24] and PARFormer [61]. What this brings is an increase in training time and inference time, which is detrimental to some UAV airborne models with limited memory. This requires further light-weighting of the model or improving the performance of the UAV.

5. Conclusions

In this paper, we present an end-to-end remote vision-based PAR model that can effectively model potential relations of regions and attributes. In contrast to existing methods, the proposed MRG-T fully explores the three relationships of inter-regions, inter-attributes, and mapping of region and attribute. Extensive experiments conducted on three publicly available datasets and a large aerial person imagery dataset demonstrate that the proposed method shows the effectiveness of mask-relation-guided modeling in the remote vision-based PAR task.

The real-world application of remote vision-based PAR in aerial imagery faces the following challenges. (1) Model lightweight issue: Current deep-learning models usually require a large amount of computing resources and are difficult to be implemented in UAVs due to limited computing power. (2) Drone battery issue: The flight duration of a drone is limited by its battery capacity. Long flights deplete the battery, restricting the drone's endurance during surveillance missions. (3) Monitoring picture quality issues: The aerial images are captured at high altitudes and may be affected by factors such as weather, lighting, and occlusion, resulting in poor image quality.

In the future, we believe that more and more researchers will invest in PAR based on remote vision. However, due to many limitations such as cost, future research can be discussed from further improving model performance or more lightweight model methods. It should be noted that everyone should pay attention to privacy protection and other aspects. We believe that in the future, there will be a more efficient, intelligent, and socially friendly pedestrian attribute recognition system based on remote vision.

Author Contributions: Conceptualization, S.Z. and Y.L.; methodology, Y.L. and X.W.; validation, X.W., Z.C. and L.L.; writing—original draft preparation, S.Z. and Y.L.; supervision, S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: The work is supported by the National Natural Science Foundation of China (no. 62271409).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, X.; Zheng, S.; Yang, R.; Zheng, A.; Chen, Z.; Tang, J.; Luo, B. Pedestrian attribute recognition: A survey. *Pattern Recognit.* **2022**, *121*, 108220. [CrossRef]
2. Schumann, A.; Stiefelhagen, R. Person re-identification by deep learning attribute-complementary information. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 20–28.
3. Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Hu, Z.; Yan, C.; Yang, Y. Improving person re-identification by attribute and identity learning. *Pattern Recognit.* **2019**, *95*, 151–161. [CrossRef]
4. Zhu, Y.; Wang, T.; Zhu, S. Adaptive Multi-Pedestrian Tracking by Multi-Sensor: Track-to-Track Fusion Using Monocular 3D Detection and MMW Radar. *Remote Sens.* **2022**, *14*, 1837. [CrossRef]
5. Zhang, S.; Huang, J.B.; Lim, J.; Gong, Y.; Wang, J.; Ahuja, N.; Yang, M.H. Tracking persons-of-interest via unsupervised representation adaptation. *Int. J. Comput. Vision* **2020**, *128*, 96–120. [CrossRef]
6. Shi, Y.; Wei, Z.; Ling, H.; Wang, Z.; Shen, J.; Li, P. Person retrieval in surveillance videos via deep attribute mining and reasoning. *IEEE Trans. Multimed.* **2020**, *23*, 4376–4387. [CrossRef]
7. Zhang, S.; Li, Y.; Mei, S. Exploring Uni-Modal Feature Learning on Entities and Relations for Remote Sensing Cross-Modal Text-Image Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–17. [CrossRef]
8. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
9. Zhao, R.; Lang, C.; Li, Z.; Liang, L.; Wei, L.; Feng, S.; Wang, T. Pedestrian attribute recognition based on attribute correlation. *Multimed. Syst.* **2022**, *28*, 1069–1081. [CrossRef]
10. Zhao, X.; Sang, L.; Ding, G.; Guo, Y.; Jin, X. Grouping attribute recognition for pedestrian with joint recurrent learning. In Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; Volume 2018, pp. 3177–3183.
11. Zhao, X.; Sang, L.; Ding, G.; Han, J.; Di, N.; Yan, C. Recurrent attention model for pedestrian attribute recognition. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 9275–9282. [CrossRef]

12. Li, Y.; Gupta, A. Beyond grids: Learning graph representations for visual recognition. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018; Volume 31.
13. Li, Q.; Zhao, X.; He, R.; Huang, K. Visual-semantic graph reasoning for pedestrian attribute recognition. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 8634–8641. [CrossRef]
14. Tan, Z.; Yang, Y.; Wan, J.; Guo, G.; Li, S.Z. Relation-aware pedestrian attribute recognition with graph convolutional networks. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12055–12062. [CrossRef]
15. Tang, Z.; Huang, J. DRFormer: Learning dual relations using Transformer for pedestrian attribute recognition. *Neurocomputing* **2022**, *497*, 159–169. [CrossRef]
16. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Online, 3–7 May 2021.
17. Hu, H.; Zhou, G.T.; Deng, Z.; Liao, Z.; Mori, G. Learning structured inference neural networks with label relations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2960–2968.
18. Wang, J.; Zhu, X.; Gong, S.; Li, W. Attribute recognition by joint recurrent learning of context and correlation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 531–540.
19. Wu, J.; Huang, Y.; Gao, Z.; Hong, Y.; Zhao, J.; Du, X. Inter-Attribute awareness for pedestrian attribute recognition. *Pattern Recognit.* **2022**, *131*, 108865. [CrossRef]
20. Li, D.; Chen, X.; Zhang, Z.; Huang, K. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In Proceedings of the IEEE International Conference on Multimedia and Expo, San Diego, CA, USA, 23–27 July 2018; pp. 1–6.
21. Liu, P.; Liu, X.; Yan, J.; Shao, J. Localization guided learning for pedestrian attribute recognition. In Proceedings of the British Machine Vision Conference, Newcastle, UK, 3–6 September 2018; p. 142.
22. Tang, C.; Sheng, L.; Zhang, Z.; Hu, X. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4997–5006.
23. Sarafianos, N.; Xu, X.; Kakadiaris, I.A. Deep imbalanced attribute classification using visual attention aggregation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 680–697.
24. Cheng, X.; Jia, M.; Wang, Q.; Zhang, J. A Simple Visual-Textual Baseline for Pedestrian Attribute Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6994–7004. [CrossRef]
25. Wu, J.; Liu, H.; Jiang, J.; Qi, M.; Ren, B.; Li, X.; Wang, Y. Person attribute recognition by sequence contextual relation learning. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 3398–3412. [CrossRef]
26. Li, T.; Liu, J.; Zhang, W.; Ni, Y.; Wang, W.; Li, Z. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 19–25 June 2021; pp. 16266–16275.
27. Cao, Y.; He, Z.; Wang, L.; Wang, W.; Yuan, Y.; Zhang, D.; Zhang, J.; Zhu, P.; Van Gool, L.; Han, J.; et al. VisDrone-DET2021: The vision meets drone object detection challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2847–2854.
28. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Fan, H.; Hu, Q.; Ling, H. Detection and tracking meet drones challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7380–7399. [CrossRef] [PubMed]
29. Zhang, S.; Zhang, Q.; Yang, Y.; Wei, X.; Wang, P.; Jiao, B.; Zhang, Y. Person re-identification in aerial imagery. *IEEE Trans. Multimedia* **2020**, *23*, 281–291. [CrossRef]
30. Zhang, S.; Yang, Q.; Cheng, D.; Xing, Y.; Liang, G.; Wang, P.; Zhang, Y. Ground-to-Aerial Person Search: Benchmark Dataset and Approach. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 789–799.
31. Akbari, Y.; Almaadeed, N.; Al-Maadeed, S.; Elharrouss, O. Applications, databases and open computer vision research from drone videos and images: A survey. *Artif. Intell. Rev.* **2021**, *54*, 3887–3938. [CrossRef]
32. Zhu, J.; Liao, S.; Lei, Z.; Yi, D.; Li, S. Pedestrian attribute classification in surveillance: Database and evaluation. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Sydney, Australia, 2–8 December 2013, pp. 331–338.
33. Deng, Y.; Luo, P.; Loy, C.C.; Tang, X. Pedestrian attribute recognition at far distance. In Proceedings of the ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 789–792.
34. Liu, X.; Zhao, H.; Tian, M.; Sheng, L.; Shao, J.; Yi, S.; Yan, J.; Wang, X. Hydraplus-net: Attentive deep features for pedestrian analysis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 350–359.
35. Tan, Z.; Yang, Y.; Wan, J.; Hang, H.; Guo, G.; Li, S.Z. Attention-based pedestrian attribute analysis. *IEEE Trans. Image Process.* **2019**, *28*, 6126–6140. [CrossRef]
36. Li, D.; Chen, X.; Huang, K. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In Proceedings of the Asian Conference on Pattern Recognition, Kuala Lumpur, Malaysia, 3–6 November 2015; pp. 111–115.
37. Sarfraz, M.S.; Schumann, A.; Wang, Y.; Stiefelhagen, R. Deep view-sensitive pedestrian attribute inference in an end-to-end model. In Proceedings of the British Machine Vision Conference, London, UK, 4–7 September 2017.

38. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 5998–6008.
39. Tenney, I.; Das, D.; Pavlick, E. BERT rediscovers the classical NLP pipeline. In Proceedings of the Conference of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 4593–4601.
40. Tetko, I.V.; Karpov, P.; Van Deursen, R.; Godin, G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nat. Commun.* **2020**, *11*, 5575. [CrossRef]
41. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
42. He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; Jiang, W. Transreid: Transformer-based Object Re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 15013–15022.
43. Ren, Z.; Gou, S.; Guo, Z.; Mao, S.; Li, R. A mask-guided transformer network with topic token for remote sensing image captioning. *Remote Sens.* **2022**, *14*, 2939. [CrossRef]
44. Reedha, R.; Dericquebourg, E.; Canals, R.; Hafiane, A. Transformer neural network for weed and crop classification of high resolution UAV images. *Remote Sens.* **2022**, *14*, 592. [CrossRef]
45. Liu, Y.; Liao, Y.; Lin, C.; Jia, Y.; Li, Z.; Yang, X. Object tracking in satellite videos based on correlation filter with multi-feature fusion and motion trajectory compensation. *Remote Sens.* **2022**, *14*, 777. [CrossRef]
46. Xu, F.; Liu, J.; Song, Y.; Sun, H.; Wang, X. Multi-exposure image fusion techniques: A comprehensive review. *Remote Sens.* **2022**, *14*, 771. [CrossRef]
47. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *ACM Comput. Surv.* **2022**, *54*, 1–41. [CrossRef]
48. Gabeur, V.; Sun, C.; Alahari, K.; Schmid, C. Multi-modal transformer for video retrieval. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 214–229.
49. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10578–10587.
50. Chen, S.; Hong, Z.; Liu, Y.; Xie, G.S.; Sun, B.; Li, H.; Peng, Q.; Lu, K.; You, X. Transzero: Attribute-guided transformer for zero-shot learning. *Proc. AAAI Conf. Artif. Intell.* **2022**, *2*, 3. [CrossRef]
51. Wu, X.; Li, Y.; Long, J.; Zhang, S.; Wan, S.; Mei, S. A remote-vision-based safety helmet and harness monitoring system based on attribute knowledge modeling. *Remote Sens.* **2023**, *15*, 347. [CrossRef]
52. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
53. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 28 March 2024).
54. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
55. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–12 December 2020; Volume 33, pp. 1877–1901.
56. Bao, H.; Dong, L.; Wei, F. Beit: Bert pre-training of image transformers. In Proceedings of the International Conference on Learning Representations, Online, 25–29 April 2022.
57. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.
58. Tao, H.; Duan, Q.; Lu, M.; Hu, Z. Learning discriminative feature representation with pixel-level supervision for forest smoke recognition. *Pattern Recognit.* **2023**, *143*, 109761. [CrossRef]
59. Lin, T.; Joe, I. An Adaptive Masked Attention Mechanism to Act on the Local Text in a Global Context for Aspect-Based Sentiment Analysis. *IEEE Access* **2023**, *11*, 43055–43066. [CrossRef]
60. Lee, G.; Cho, J. STDP-Net: Improved Pedestrian Attribute Recognition Using Swin Transformer and Semantic Self-Attention. *IEEE Access* **2022**, *10*, 82656–82667. [CrossRef]
61. Fan, X.; Zhang, Y.; Lu, Y.; Wang, H. PARFormer: Transformer-based Multi-Task Network for Pedestrian Attribute Recognition. *IEEE Trans. Circ. Syst. Video Technol.* **2024**, *33*, 411–423. [CrossRef]
62. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
63. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
64. Lanchantin, J.; Wang, T.; Ordonez, V.; Qi, Y. General multi-label image classification with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16478–16488.

65. Li, D.; Zhang, Z.; Chen, X.; Huang, K. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE Trans. Image Process.* **2018**, *28*, 1575–1590. [CrossRef]
66. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
67. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
68. Liu, Z.; Zhang, Z.; Li, D.; Zhang, P.; Shan, C. Dual-branch self-attention network for pedestrian attribute recognition. *Pattern Recognit. Lett.* **2022**, *163*, 112–120. [CrossRef]
69. Jia, J.; Huang, H.; Chen, X.; Huang, K. Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting. *arXiv* **2021**, arXiv:2107.03576.
70. Jia, J.; Chen, X.; Huang, K. Spatial and semantic consistency regularizations for pedestrian attribute recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 962–971.
71. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.