C^{2}T-Net: Channel-Aware Cross-Fused Transformer-Style Networks for Pedestrian Attribute Recognition

Doanh C. Bui[1] Thinh V. Le[2] Ba Hung Ngo[3]
[1]School of Electrical Engineering, Korea University, Republic of Korea
[2]University of Information Technology, Vietnam National University, Vietnam
[3]Graduate School of Data Science, Chonnam National University, Republic of Korea
doanhbc@korea.ac.kr, 20520781@gm.uit.edu.vn, ngohung@jnu.ac.kr

Abstract

Pedestrian attribute recognition (PAR) poses a significant challenge but holds practical significance in various security applications, including surveillance. In the scope of the UPAR challenge, this paper introduces the Channel-Aware Cross-Fused Transformer-Style Networks (C^{2}T-Net). This network effectively integrates two powerful transformer-style networks, namely the Swin Transformer (SwinT) and a customized variant of the vanilla vision transformer (EVA ViT). The aim is to capture both local and global aspects of an individual for precise attribute recognition. To facilitate the understanding of intricate relationships among channels, a channel-aware self-attention mechanism is devised and integrated into each SwinT block. Furthermore, the fusion of features from the two transformer-style networks is accomplished through cross-fusion, enabling each network to mutually amplify and boost the textural nuances present in the other. The efficacy of the proposed model has been demonstrated through its performance on three PAR benchmarks: PA100K, PETA, and the UPAR2024 private test. With respect to the PA100K benchmark, our approach has achieved state-of-the-art results when compared to models that do not employ any pre-training techniques. Our performance on the PETA dataset remains competitive, standing on par with other cutting-edge models. Notably, our model achieved runner-up performance on the UPAR2024-track-1 test set. Source code is available at https://github.com/caodoanh2001/upar_challenge.

1. Introduction

Pedestrian analysis has emerged as a critical area of focus in camera surveillance and pedestrian characteristic research. Within this domain, the recognition of pedestrian attributes (PAR) stands out as a pivotal sub-problem. This problem, characterized by its formulation as a multi-label classification task, necessitates the ability to predict multiple categories within a single prediction. Notably, recent studies [15, 18, 22, 25, 26, 32] have harnessed deep-learning-based models to yield promising results across various datasets such as PA100K [18], PETA [7], and RAPv2 [17], highlighting the significance of this area. However, the continuous influx of samples from diverse datasets has underscored the pressing need for the development of more robust methodologies [6, 24].

In this context, the exploration of feature representations within the PAR problem domain has remained a top priority, as evidenced by the global-to-local aspect highlighted in previous research [18]. Besides, hybrid models, which involve the combination of multiple networks, including convolutional neural networks (CNNs) or transformer-style

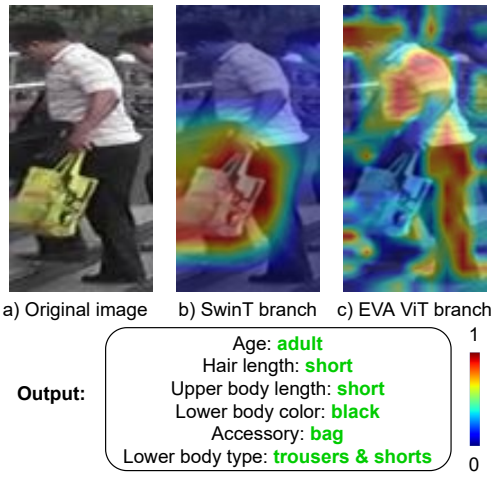

Figure 1. **Prediction for a sample from UPAR2024 dev-test of our C^{2}T-Net.** Focus visualization from two network branches using GradCAM [23]. **Green** text denotes correctly predicted attributes compared to ground-truth.

models, have demonstrated their ability to provide robust representations, thereby empowering the head classifiers to perform with superior efficacy [8,20,30]. Against the backdrop of the UPAR challenge 2024 [5], where pedestrian attribute recognition takes center stage, the development of a meticulously designed hybrid model integrating a fusion mechanism holds the promise of delivering even more remarkable outcomes. Leveraging the proven effectiveness of transformer-style models, our proposition aims to capitalize on their capabilities to craft a hybrid model that can yield robust feature representations, thereby paving the way for substantial advancements within the PAR problem domain.

In the pursuit of enhancing pedestrian attribute recognition, delving into local information assumes pivotal significance. The different regions of the body, including the upper, middle, and lower sections, often serve as the bearers of crucial visual characteristics. Our research endeavors have shed light on the suitability of Swin Transformer (SwinT) [19] for investigating these facets. Central to SwinT's efficacy is the shifted window multi-head self-attention (W-MSA) mechanism, which, in contrast to traditional self-attention (SA), computes attention weights exclusively within designated windows, thereby enabling window shifting across blocks. This unique approach facilitates the assimilation of localized information while nurturing inter-region awareness across successive layers.

While the emphasis on local body parts is essential, it is imperative to strike a balance to prevent the incorporation of extraneous and noisy information, particularly in the presence of complex backgrounds within the images. The shifted window of SwinT, in some instances, might inadvertently focus on irrelevant regions, thus necessitating the integration of a fully-connected self-attention mechanism similar to traditional self-attention. This mechanism, inspired by the Vision Transformer model (ViT) [10], operates across all image patches, effectively filtering out superfluous background elements.

Herein, we propose **C**hannel-Aware **C**ross-Fused **T**ransformer-Style **Net**works (C^2T-Net), harnessing the strengths of both SwinT and the traditional self-attention mechanism from ViT. In addition, the quality of features obtained from the SwinT branch is enhanced through a module called the channel-aware self-attention mechanism (CASA), improving the flow of information within the channel perspective. Additionally, a cross-fusion (CF) module has been designed to promote mutual awareness among the final feature vectors of each branch, yielding a fused final vector well-suited for the multi-label classification task inherent in pedestrian attribute recognition. In summary, our contributions are listed below:

1. We introduce an advanced approach that leverages the capabilities of two cutting-edge transformer-style networks: SwinT and EVA ViT models.

2. We design a channel-aware attention mechanism that operates atop each SwinT block, facilitating the comprehensive extraction of highlighted features from each position within the SwinT feature maps.

3. We propose the implementation of a cross-fusion mechanism, resulting in the creation of fused feature vectors tailored for pedestrian attribute recognition. This mechanism effectively integrates the valuable information derived from the ViT patch tokens and the feature maps originating from the SwinT network, ensuring the optimal utilization of pertinent data for enhanced performance.

4. The proposed approach is evaluated on the PA100K [18] and PETA [7] benchmarks, alongside the private test set of the UPAR challenge 2024 track 1 (UPAR2024-track-1 test set). In detail, our method achieves a new state-of-the-art performance on the PA100K dataset, outperforming models that did not undergo any pre-training tasks. In the case of the PETA dataset, we attain competitive results comparable to other state-of-the-art methods. Lastly, our approach secures a 2^{nd} ranking on the UPAR2024-track-1 test set.

2. Related Work

2.1. PAR-specific design models

Imbalance-aware techniques. In surveillance contexts, the performance of PAR models is often hampered by the unbalanced distribution of human attributes. To tackle this issue, Li et al. implemented a weighted binary cross-entropy loss function [15] and a technique that involves duplicating images at random, thereby equalizing the number of positive and negative examples in the training dataset, effectively addressing the problem of imbalanced data distribution.

Attention-based mechanisms. Some researchers [18, 22] exploited the utilization of the visual attention mechanism in attribute recognition. Sarafianos et al. introduced an effective approach [22] to aggregate visual attention masks at various scales to enhance the learning process of uncertain samples and maintain the local class structures when dealing with imbalanced data. Liu et al. applied the concept of multi-level fusion through the use of multi-scale attentive maps [18] generated by using Visual semantic attributes as a mid-level feature, thereby enriching the final feature representation. Considering attributes exhibit notable spatial correlations with human structures, Li et al. [16] attempted to explore the effectiveness of pose information in the task of pedestrian attribute recognition. To achieve the final attribute prediction, Li et al. employed the Spatial Transformer Network (STN) [13] to merge key points with the

global body-based outputs.

Multi-scale features aggregation. Tan et al. proposed an end-to-end unified model [25] that incorporates two GCN-based modules for capturing both attribute and contextual relations, the final prediction is made by leveraging the aggregated features from these two branches. Tang et al. presented an end-to-end framework [26] consisting of an ALM and FPN module, designed to carry out attribute-specific localization across multiple scales in order to identify the most distinctive attribute regions. To address the drop in performance when encountering individuals at a distance, Zhong et al. proposed the MSSC [32]. This module enhances features for inconspicuous attributes by aggregating contextual information across receptive fields, integrating low to high-level features using non-local attention, and establishing long-range dependencies in pyramid feature maps across spatial scales.

Simple strong baselines. Considering PAR as a multi-label classification problem involving binary attributes as in [14], Specker et al. introduced a simple classification framework [24] that consists of a ConvNeXt backbone and a fully-connected classification head with a final layer using Sigmoid activation. Additionally, several enhancements and techniques were applied to the baseline, including exponential moving averages of model weights, suitable batch sizes, label smoothing, dropout, and data augmentation methods.

2.2. Transformer-style models

The Transformer architecture was initially developed for machine translation tasks in the field of natural language processing (NLP). Its inherent ability to effectively capture long-range dependencies in data through global self-attention mechanisms quickly led to Transformer-based frameworks dominating this domain. Inspired by this accomplishment, Alexey et al. introduced the pioneering Vision Transformer (ViT) [9] for image classification, achieving impressive results compared to state-of-the-art CNNs. ViT's success has subsequently led to research efforts aimed at enhancing its performance.

One notable contribution, Deit [28], focuses on various training strategies designed to mitigate the challenges associated with pre-training on large datasets. Swin Transformer [19], on the other hand, takes inspiration from the inductive biases of locality, hierarchy, and translation invariance. It introduced a shifted windows mechanism, enabling it to serve in a wide range of image recognition tasks. In addition, EVA [11] distills the multi-modal knowledge to scale up ViT by leveraging unlabeled images with the large-scale pre-trained image-text model CLIP [21] and achieves remarkable results in various vision downstream tasks including image recognition, video action recognition, object detection, instance segmentation, and semantic segmentation, all without the need for extensive supervised training.

3. Methodology

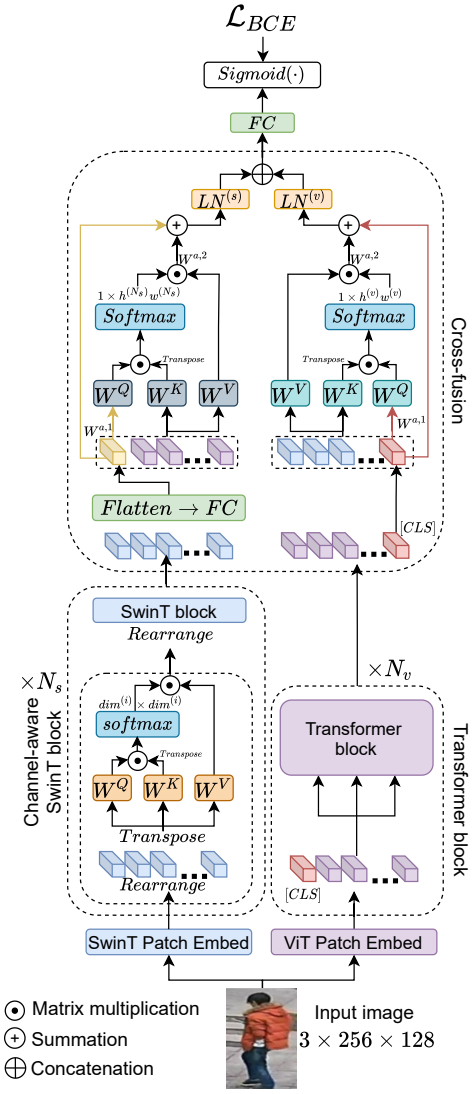

Figure 2. Overview of our proposed approach: Two network branches are designed, namely SwinT [19] and Eva vanilla ViT [11]. At the top of each SwinT block, there is channel-aware attention that calculates attention weights by utilizing spatial information. At the core of our network, a cross-fusion mechanism is employed to fully leverage patch tokens from the ViT branch and feature maps from the SwinT branch, enabling each final feature vector to become aware of the other. Finally, the mean is taken across channels, yielding discrete probabilities for person attribute recognition.

Overall architecture. Our overall architecture (See Fig. 2) comprises two transformer-style models: SwinT and an EVA-based Vanilla ViT, both processing images at a resolution of 256×128. The `base` version of SwinT is em-

ployed, with $embed\_dim = 1024$, and consists of four SwinT blocks. To enhance channel awareness in token features, we introduced channel-aware self-attention (CASA) after applying shifted window multi-head self-attention operations on each SwinT block. Considering an input image $I \in \mathbb{R}^{3\times H\times W}$, the process can be formulated as follows:

$$z_s^{(0)} = \text{PatchEmbed}(I), \tag{1}$$

$$z_s^{(i)} = \text{SB}^{(i)}\Big(\text{CASA}^{(i)}\big(z_s^{(i-1)}\big)\Big), \quad 1 \le i \le N_s, \tag{2}$$

where $z_s^{(0)} \in \mathbb{R}^{dim\times H_s\times W_s}$ $(H_s = H/4, W_s = W/4)$ represents the embedded tokens from the original image I. Here, CASA$(\cdot)$ refers to the channel-aware self-attention, detailed in the following sub-sections, and SB$^{(i)}(\cdot)$ denotes the i^{th} SwinT block introduced in the study [19], with N_s as the total number of SwinT blocks. Following N_s SwinT blocks, we obtain final feature maps $z_s^{(N_s)} \in \mathbb{R}^{dim\times h_s\times w_s}$, where $h_s = H/32$ and $w_s = W/32$. These $dim \times h_s \times w_s$ features are then pooled into dim features using average pooling:

$$\mathbf{z}^s = \text{AvgPooling}(z_s^{(N_s)}), \tag{3}$$

where $\mathbf{z}^s \in \mathbb{R}^{dim}$ denotes the pooled features from the SwinT branch. However, note that $z_s^{(N_s)}$ is still utilized for cross-fusion.

Simultaneously, we incorporate another vanilla EVA-based ViT branch (`large` version, $embed\_dim = 1024$) consisting of N_v transformer blocks. Unlike the original ViT, where the patch size commonly has the same values for width w_v and height h_v, we utilize different values to better suit the shape of the person sample. Processing the image I using ViT results in a sequence of patch tokens $z_v^{(N_v)} \in \mathbb{R}^{dim\times(h_v w_v)}$ along with the `[CLS]` token, denoted as $\mathbf{z}^v \in \mathbb{R}^{dim}$ for convenience.

We then perform cross-fusion to enable the two branches of networks to be aware of each other in an attention-style manner:

$$\mathbf{z}^{s'} = \text{SVCF}(\mathbf{z}^s, z_v^{(N_v)}), \tag{4}$$

$$\mathbf{z}^{v'} = \text{VSCF}(\mathbf{z}^v, z_s^{(N_s)}), \tag{5}$$

where SVCF$(\cdot)$ and VSCF$(\cdot)$ refer to SwinT-ViT and ViT-SwinT cross-fusions, respectively. $\mathbf{z}^{s'}$ and $\mathbf{z}^{v'}$ represent cross-aware feature vectors, both with dimensions of dim.

Finally, $\mathbf{z}^{s'}$ and $\mathbf{z}^{v'}$ are applied two independent layer normalizations, to capture pattern-specific distributions, then concatenated and passed into the fully connected layer to map dim to n_{att} (the number of attributes that need to be recognized).

Channel-aware Self-attention (CASA). The spirit of transformer-style models is the (multi-head) self-attention mechanism [29], which can construct connections among all patch tokens in the learning process. Given a sequence including N tokens $\mathbf{z} = \{z_i\}_{i=1}^N$, three learned matrices $\mathbf{W}^q$, $\mathbf{W}^k$, and $\mathbf{W}^v$ are multiplied with $\mathbf{z}$ to yield three query ($\mathbf{q} \in \mathbb{R}^{N\times\text{dim}}$), key ($\mathbf{k} \in \mathbb{R}^{N\times\text{dim}}$), and value ($\mathbf{v} \in \mathbb{R}^{N\times\text{dim}}$) tokens, respectively. Subsequently, the attention weight matrix is computed using the scaled dot product between $\mathbf{q}$ and $\mathbf{v}$, followed by the softmax function: $att = \text{softmax}(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d_k}})$. Finally, the resulting att is multiplied with $\mathbf{v}$ to emphasize the importance of specific tokens. Despite its capabilities in computing attention weights based on spatial aspects, self-attention, unfortunately, does not inherently account for channel relationships. Consequently, the channel aspect is often overlooked. To address this limitation, we introduce the concept of channel-aware self-attention at the top of each SwinT block. Given output $z_s^i \in \mathbb{R}^{dim^{(i)}\times h^{(i)}\times w^{(i)}}$ at i^{th} SwinT block, this approach is formulated as CASA$(\cdot)$ operation, which include following steps:

$$\begin{aligned}
&z_s^{i,T} = \text{Transpose}(\text{Group}(z_s^i)), z_s^{i,T} \in \mathbb{R}^{(h^{(i)}w^{(i)})\times dim},\\
&\mathbf{q}^i = z_s^{i,T} \odot \mathbf{W}^{q,i}, \mathbf{k}^i = z_s^{i,T} \odot \mathbf{W}^{k,i}, \mathbf{v}^i = z_s^{i,T} \odot \mathbf{W}^{v,i},\\
&\text{where } \mathbf{W}^{\cdot,i} \in \mathbb{R}^{h^{(i)}w^{(i)}\times h^{(i)}w^{(i)}},\\
&att = \text{softmax}\big(\tfrac{\mathbf{q}^i\mathbf{k}^{iT}}{\sqrt{d_k}}\big), att \in \mathbb{R}^{h^{(i)}w^{(i)}\times h^{(i)}w^{(i)}},\\
&z_s^{i,sb,T} = z_s^{i,T} + att \odot z_s^{i,T},\\
&z_s^{i,sb} = \text{Ungroup}(\text{Transpose}(z_s^{i,sb,T})),\\
&z_s^{i,sb} \in \mathbb{R}^{dim\times h^{(i)}\times w^{(i)}},\\
&z_s^{i+1} = \text{SB}^{(i)}(z_s^{i,sb}),
\end{aligned} \tag{6}$$

where $z_s^{i,T}$ denotes the transpose matrix of Group(z_s^i), with the group of spatial resolutions $(h^{(i)} \times w^{(i)})$ considered as rows of the matrix. Subsequently, we compute the channel-aware attention weight matrix att to identify crucial features, based on the spatial tokens. Finally, att is multiplied with the transpose matrix $z_s^{i,T}$ along with the skip connection, resulting in $z_s^{i,sb,T}$. This is then rearranged to $z_s^{i,sb}$, which represents the input to the i^{th} SwinT block (SB$^{(i)}(\cdot)$).

Cross-fusion (CF). At the head of the two network branches, we obtained feature vectors $\mathbf{z}^s \in \mathbb{R}^{dim}$ and $\mathbf{z}^v \in \mathbb{R}^{dim}$, as well as feature maps $z_s^{(N_s)} \in \mathbb{R}^{dim\times h^{(N_s)}\times w^{(N_s)}}$ and $z_v^{(N_v)} \in \mathbb{R}^{(h_v\times w_v)\times dim}$. These represent the SwinT feature vector, the feature vector of ViT `[CLS]` token, the feature maps from SwinT, and the sequence of ViT tokens, respectively. Simply using the concatenate of $\mathbf{z}^s$ and $\mathbf{z}^v$ also show the effectiveness. However, it does not fully take advantage of $z_v^{(N_v)}$ and $z_s^{(N_s)}$, which also contain useful feature values. Herein, given $\mathbf{z}^s$, $z_v^{(N_v)}$ that are considered to perform cross-fusion, inspired by [1], we design the cross-

attention as follows:

$$\begin{aligned}
&\mathbf{z}^{s,a} = \mathbf{z}^{s} \odot \mathbf{W}^{a,1}, \mathbf{W}^{a,1} \in \mathbb{R}^{dim \times dim} \\
&\mathbf{z}^{s,c} = \text{Concatenate}(\mathbf{z}^{s,a}, z_v^{(N_v)}), \mathbf{z}^{s,c} \in \mathbb{R}^{(h_v w_v + 1) \times dim} \\
&\mathbf{q} = \mathbf{z}^{s} \odot \mathbf{W}^{q}, \mathbf{k} = \mathbf{z}^{s,c} \odot \mathbf{W}^{k}, \mathbf{v} = \mathbf{z}^{s,c} \odot \mathbf{W}^{v}, \\
&att = \text{softmax}(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d_k}}), att \in \mathbb{R}^{1 \times (h_v w_v + 1)}, \\
&\mathbf{z}^{s'} = (\mathbf{z}^{s,c} + \mathbf{z}^{s,c} \odot att) \odot \mathbf{W}^{a,2}, \mathbf{W}^{a,2} \in \mathbb{R}^{dim \times dim},
\end{aligned} \tag{7}$$

where $\mathbf{W}^{a,1}$ and $\mathbf{W}^{a,2}$ are the projection matrices for dimension alignment. $\mathbf{z}^{s,a}$ denotes the aligned token features, and $\mathbf{z}^{s,c}$ represents the concatenation of $\mathbf{z}^{s,a}$ and $z_v^{(N_v)}$. By employing this approach, it becomes evident that a feature vector from the SwinT, i.e., a single token, can compute self-attention with a sequence of tokens from EVA ViT, resulting in attention weights $att \in \mathbb{R}^{1 \times (h_v w_v + 1)}$. In other words, this method enables a token to become aware of other sequences of tokens and compresses the knowledge across channels. Hence, it explains how $\mathbf{z}^{s}$ becomes aware of $z_v^{(N_v)}$. This operation is referred to as the SVCF$(\cdot)$ operation. Regarding how $\mathbf{z}^{v}$ becomes aware of $z_s^{(N_s)}$, the processing steps in Eq. 9 are also performed and considered as the VSCF$(\cdot)$ operation. However, it is worth noting that $z_s^{(N_s)}$ still remains in the form of a tensor, i.e., $z_s^{(N_s)} \in \mathbb{R}^{dim \times h^{(N_s)} \times w^{(N_s)}}$. Consequently, we first need to group the spatial resolutions of $z_s^{(N_s)}$ together, treating them as a sequence of tokens, to perform cross-fusion with $\mathbf{z}^{v}$ through a cross-attention mechanism.

Independent Layer Normalization. With two transformer-style networks offering distinct perspectives, we introduce two independent layer normalizations before fusion. These normalizations are tailored to capture two unique distributions for the representations $\mathbf{z}^{s'}$ and $\mathbf{z}^{v'}$. This can be expressed as the following equations:

$$\begin{aligned}
\mathbf{M}^{s} &= \text{LN}^{(s)}(\mathbf{z}^{s'}; \alpha_s, \beta_s), \\
\mathbf{M}^{v} &= \text{LN}^{(v)}(\mathbf{z}^{v'}; \alpha_v, \beta_v),
\end{aligned} \tag{8}$$

where $\text{LN}^{(s)}$ and $\text{LN}^{(v)}$ denote two layer normalizations applied to the feature vectors from SwinT ($\mathbf{z}^{s'}$) and EVA ViT ($\mathbf{z}^{v'}$), respectively. The learnable scale and shift parameters, α_s and β_s, are applied to the affine feature values of $\mathbf{z}^{s'}$, while α_v and β_v perform the same function for $\mathbf{z}^{v'}$.

Subsequently, we concatenate the two normalized representations and pass them through a fully-connected layer to obtain N_{attr}-dimensional logits. Finally, we apply the Sigmoid$(\cdot)$ function to derive the final output for multi-label classification:

$$\begin{aligned}
\mathbf{M}^{sv} &= \text{FC}(\text{Concatenate}(\mathbf{M}^{s}, \mathbf{M}^{v})), \\
\hat{\mathbf{y}} &= \text{Sigmoid}(\mathbf{M}^{sv}).
\end{aligned} \tag{9}$$

Loss function. To facilitate multi-label classification, we employ the binary cross-entropy (BCE) loss function for supervised learning. Considering $\hat{\mathbf{y}} \in \mathbb{R}^{N_{attr}}$ as the output from the Sigmoid$(\cdot)$ function and $\mathbf{y} = \{y_i\}_{i=1}^{N_{attr}} \in \mathbb{R}^{N_{attr}}$ as the one-hot encoding ground-truth, where $y_i \in \{0, 1\}$, the binary cross-entropy loss function is defined as follows:

$$\mathcal{L}_{BCE}(\hat{\mathbf{y}}, \mathbf{y}) = -\left(\mathbf{y} \log(\hat{\mathbf{y}}) + (1 - \mathbf{y}) \log(1 - \hat{\mathbf{y}})\right). \tag{10}$$

4. Challenge Description

The Pedestrian Attribute Recognition and Person Retrieval Challenge at WACV 2024 [5] is split into two tracks that share the same data sources consisting of three public benchmark datasets PA100K [18], PETA [7]. Track 1 points to the "**Pedestrian Attribute Recognition**" task to classify the semantic attributes of persons under domain shifts. Track 2 is designed for the "**Attribute-based Person Retrieval**" task, where the methods are proposed aiming to match the attribute-based person retrieval to a specific attribute description. The attributes of these datasets are listed in Tab. 1. In this paper, we focus on the task 1.

5. Results

Datasets. It is noteworthy that, besides reporting results for the UPAR2024 dataset, we conducted experiments on two standard benchmarks for the PAR problem: PA100K [18] and PETA [7] datasets, both of which are included in it. The statistics of these datasets are presented in Tab. 2. The PA100K dataset comprises 100,000 pedestrian images from 598 scenes with 26 attributes, collected from various camera settings with different lighting conditions, image resolutions, and environments. This dataset includes multiple object-level attributes such as a handbag, phone, upper-clothing, and global attributes like gender, age, and *etc.* The PETA dataset is created by aggregating 19,000 pedestrian images from 10 publicly available datasets, covering a variety of indoor and outdoor scenes. The dataset is annotated with 61 binary attributes and four multiclass attributes. The UPAR2024 dataset provided by the challenge organizer includes samples from PA100K, PETA, and Market1501 [31] datasets, which are used for training and validation (UPAR2024 dev-test), and the testing set (UPAR2024 private-test) is collected from a private source.

Implemental details. As previously stated, we have integrated two backbone networks into our system: SwinT and EVA ViT. The SwinT architecture employs the `base` version, while for EVA ViT, we have implemented the `large` version. Notably, both networks are configured with an $embed\_dim$ of 1024, and solely rely on pre-trained weights from ImageNet as their initial weights. We completed training the entire network within 10 epochs, observing that convergence typically occurs around the 4th or

Category	Age	Gender	Hair length	UB clothing length	UB clothing color	LB clothing length	LB clothing color	LB clothing type	Backpack	Bag	Glasses	Hat
Attributes	Young Adult Elderly	Female	Short Long Bald	Short	Black Blue Brown Green Grey Orange Pink Purple Red White Yellow Other	Short	Black Blue Brown Green Grey Orange Pink Purple Red White Yellow Other	Trousers&Shorts Skirt&Dress	Backpack	Bag	Normal Sun	Hat

Table 1. Lists of Pedestrian Attributes in UPAR dataset [24].

	PETA [7]	PA100K [18]	UPAR2024 [5]
# scene	-	598	-
# sample	19,000	100,000	159,171
# attribute	61 (+4)	26	40
# tracklet	-	18,206	-
# resolution	from 17×39 to 169×365	from 50×100 to 758×454	from 16×43 to 338×766

Table 2. Dataset description.

5th epoch, yielding the best results. *Adam* optimizes the model, while the learning rate is dynamically adjusted using the *plateau* scheduler (with parameters: $factor = 0.1$, $patience = 4$). The initial learning rate is established at $1e-6$. During training, a batch size of 64 is employed. To accommodate the human form, all images are resized to 256×128.

Evaluation metrics. Following the UPAR Challenge at WACV'24[1], we used the harmonic mean from mA and instance-based F1 as evaluation metrics for the task 1. The mA metric is used to calculate individual attributes, while the instance-based F1 score is used to estimate the quality predictions for all attributes associated with the persons.

Main results. Tab. 3 displays a comparison between our approach and the most recent methods applied to the PA100K and PETA datasets. When it comes to the PA100K dataset, our evaluation results indicate that our proposed approach outperforms the currently employed methods, including both those with CNN-based and transformer-based backbones, in terms of mA and F1 scores. Our method surpasses state-of-the-art detectors in both two approaches with an mA score, higher than them by margins of 2.4 and 0.3, respectively. Additionally, we achieved competitive results on the PETA dataset, securing the second-highest F1 score.

Moreover, the outcomes reveal the impressive capabilities of methods employing transformer-based backbones, with CNN-based approaches consistently yielding lower results in terms of mA score. This superior performance can be attributed in part to the use of pre-training techniques in methods [3], [4], and [27]. Nevertheless, even without leveraging information from pretraining tasks, our approach still manages to deliver a competitive performance.

In order to solidify the demonstration of our method's effectiveness, we show the results of the private-test set of UPAR challenge 2024 track 1, as depicted in Tab. 4. Our method takes the second position on the leaderboard, achieving an average score of 71.74, surpassing the official baseline [5] by 2.32. It's worth noting that our method attains the highest performance in both F1_{label} and F1_{inst} scores, with an F1_{label} score that exceeds the solutions presented by sophere001 by 1.3, and an F1_{inst} score that outperforms the solutions provided by fanttee by 0.53.

Fig. 1 depicts a prediction from a sample in the UPAR2024 dev-test dataset. The EVA ViT network demonstrates its ability to capture global patterns, aiding in the accurate recognition of visual attributes. Furthermore, the SwinT blocks effectively capture local regions, correctly identifying the person's bag.

Ablations study. We conducted an ablation study using five different settings: 1) $f^{(v)}$: single EVA ViT; 2) $f^{(s)}$: single SwinT; 3) $f^{(s),CASA}$: single SwinT with CASA incorporated on top of each block; 4) $f^{(v)} \oplus f^{(s),CASA}$: the feature vectors of $f^{(v)}$ and $f^{(s),CASA}$ are concatenated; and 5) $f^{(v)} \chi f^{(s),CASA}$, denoted as C^{2}T-Net. When comparing the performance of the two individual models, $f^{(v)}$ and $f^{(s)}$, across all evaluation metrics in the PA100K and UPAR2024 datasets, our findings suggest that EVA's ability to capture global information provides an advantage in multi-class classification tasks compared to SwinT. This is because SwinT tends to focus on localized features, potentially leading to a limited focus on specific regions and overlooking valuable information from other parts of the image. Such oversight can be crucial in achieving higher accuracy across different labels. By utilizing the channel aspect information through the CASA mechanism positioned at the top of each SwinT block ($f^{(s),CASA}$), we observed an enhancement in the performance of SwinT. In Table 5, we note a minor increase of +0.7 in mA, although there was a slight drop in F1. When simply concatenating $f^{(s),CASA}$ and $f^{(v)}$, some improvements were observed, although they were marginal and not entirely clear. However, we found

[1] https://chalearnlap.cvc.uab.cat/challenge/57/description/

Method	Backbone	FT	PA100K		PETA	
			mA	F1	mA	F1
MsVAA [22]	ResNet-101	✗	–	–	84.6	86.5
VAC [12]	ResNet-50	✗	79.0	86.8	83.6	86.2
ALM [26]	BN-Inception	✗	80.7	86.5	83.6	86.9
JLAC [25]	ResNet-50	✗	82.3	87.6	87.0	87.5
VFA [2]	ResNet-50	✗	81.3	87.0	86.5	87.3
MSCC [32]	ResNet-50	✗	82.1	86.8	–	–
SB [14]	ResNet-50	✗	81.6	88.1	84.0	86.3
UPAR [24]	ResNet-50	✗	82.2	88.5	87.1	87.7
UPAR [24]	ConvNeXt-B	✗	84.8	90.2	88.4	**89.9**
SOLDIER [3]	SwinT-B	✓	86.4	–	–	–
UniHCP [4]	Enc-Dec	✓	86.2	–	–	–
PATH [27]	ViT-B	✓	86.9	–	**89.8**	–
C^{2}T-Net (ours)	SwinT-B + EVA-L	✗	**87.2**	**91.0**	88.0	89.1

Table 3. Comparison of our C^{2}T-Net with other state-of-the-art models on the PA100K and PETA datasets.

Method	Avg	mA	F1$_{label}$	F1$_{inst}$
1^{st} fanttec	**71.83**	**71.14**	45.25	72.54
3^{rd} sophere001	71.59	70.89	46.30	72.31
Official baseline [5]	69.42	67.97	43.22	70.94
C^{2}T-Net (Ours)	71.74	70.46	**47.60**	**73.07**

Table 4. Comparison of our C^{2}T-Net with other participants on the private-test set of UPAR challenge 2024 track 1.

Method	PA100K			UPAR2024 dev-test		
	Avg	mA	F1	Avg	mA	F1
$f^{(v)}$	87.9	85.2	90.7	87.6	85.8	89.4
$f^{(s)}$	85.7	82.6	88.8	86.3	84.3	88.2
$f^{(s),CASA}$	85.9	83.3	88.6	86.5	84.9	88.1
$f^{(v)} \oplus f^{(s),CASA}$	87.8	85.4	90.3	87.7	85.6	89.8
$f^{(v)} \chi f^{(s),CASA}$	**89.1**	**87.2**	**91.0**	**87.9**	**85.9**	**90.0**

Table 5. Ablation study on PA100K test set and UPAR2024 dev-test.

significant improvements across all metrics and datasets by employing the cross-fusion method before concatenation. This approach yielded the best results of 89.1% and 87.9% for the PA100K and UPAR2024 dev-test, respectively. We employed this proposed model to evaluate the performance on the UPAR2024 private-test, where it secured the 2nd rank.

6. Conclusion

In summary, this paper introduces a Channel-Aware Cross-Fused Transformer-Style Networks (C^{2}T-Net) designed to enhance the accuracy of the pedestrian attribute recognition task. C^{2}T-Net utilizes two transformer-based architectures, SwinT and EVA ViT, with one focused on capturing local features and the other on global features. Channel-aware self-attention is proposed to address the challenge of comprehending intricate relationships in the perspective of channels. Lastly, our model combines features from these two transformer-style networks to facilitate mutual understanding between local and global features, ultimately boosting the model's performance.

References

[1] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 4

[2] Ming Chen, Guijin Wang, Jing-Hao Xue, Zijian Ding, and Li Sun. Enhance via decoupling: Improving multi-label classifiers with variational feature augmentation. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1329–1333. Institute of Electrical and Electronics Engineers (IEEE), 2021. 7

[3] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15050–15061, 2023. 6, 7

[4] Yuanzheng Ci, Yizhou Wang, Meilin Chen, Shixiang Tang, Lei Bai, Feng Zhu, Rui Zhao, Fengwei Yu, Donglian Qi, and Wanli Ouyang. Unihcp: A unified model for human-centric perceptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17840–17852, 2023. 6, 7

[5] Mickael Cormier, Andreas Specker, Julio C. S. Jacques, Lennart Moritz, Jürgen Metzler, Thomas B. Moeslund, Kamal Nasrollahi, Sergio Escalera, and Jürgen Beyerer. Upar challenge 2024: Pedestrian attribute recognition and attribute-based person retrieval - dataset, design, and results. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2024. 2, 5, 6, 7

[6] Mickael Cormier, Andreas Specker, Julio Junior, CS Jacques, Lucas Florin, Jürgen Metzler, Thomas B Moeslund, Kamal Nasrollahi, Sergio Escalera, and Jürgen Beyerer. Upar challenge: Pedestrian attribute recognition and attribute-based person retrieval–dataset, design, and results. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 166–175, 2023. 1

[7] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 789–792, 2014. 1, 2, 5, 6

[8] Youcef Djenouri and Ahmed Nabil Belbachir. A hybrid visual transformer for efficient deep human activity recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 721–730, 2023. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2010. 3

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[11] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 3

[12] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 729–739, 2019. 7

[13] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks, 2016. 2

[14] Jian Jia, Houjing Huang, Xiaotang Chen, and Kaiqi Huang. Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting. *arXiv preprint arXiv:2107.03576*, 2021. 3, 7

[15] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 111–115. IEEE, 2015. 1, 2

[16] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *2018 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2018. 2

[17] Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE transactions on image processing*, 28(4):1575–1590, 2018. 1

[18] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 350–359, 2017. 1, 2, 5, 6

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 3, 4

[20] Zhenyu Liu, Zhang Zhang, Da Li, Peng Zhang, and Caifeng Shan. Dual-branch self-attention network for pedestrian attribute recognition. *Pattern Recognition Letters*, 163:112–120, 2022. 2

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[22] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 680–697, 2018. 1, 2, 7

[23] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1

[24] Andreas Specker, Mickael Cormier, and Jürgen Beyerer. Upar: Unified pedestrian attribute recognition and person retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 981–990, 2023. 1, 3, 6, 7

[25] Zichang Tan, Yang Yang, Jun Wan, Guodong Guo, and Stan Z Li. Relation-aware pedestrian attribute recognition with graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12055–12062, 2020. 1, 3, 7

[26] Chufeng Tang, Lu Sheng, Zhaoxiang Zhang, and Xiaolin Hu. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4997–5006, 2019. 1, 3, 7

[27] Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, et al. Humanbench: Towards general human-centric perception with projector assisted pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21970–21982, 2023. 6, 7

[28] Hugo Touvron, M Cord, M Douze, F Massa, A Sablayrolles, and H Jégou. Training data-efficient image transformers and distillation through attention (2020). doi: 10.48550. *arxiv*, 2012. 3

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[30] Ngoc Tu Vu, Van Thong Huynh, Trong Nghia Nguyen, and Soo-Hyung Kim. Ensemble spatial and temporal vision transformer for action units detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5769–5775, 2023. 2

[31] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 5

[32] Jiabao Zhong, Hezhe Qiao, Lin Chen, Mingsheng Shang, and Qun Liu. Improving pedestrian attribute recognition with multi-scale spatial calibration. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 1, 3, 7