Dynamic Token Selective Transformer for Aerial-Ground Person Re-Identification

Yuhai Wang[1*], Maryam Pishgar[1]
[1]University of Southern California
[*]Corresponding author
https://yuhaiw.github.io/DTS-AGPReID/

***Abstract*—Aerial-Ground Person Re-identification (AGPReID) holds significant practical value but faces unique challenges due to pronounced variations in viewing angles, lighting conditions, and background interference. Traditional methods, often involving a global analysis of the entire image, frequently lead to inefficiencies and susceptibility to irrelevant data. In this paper, we propose a novel Dynamic Token Selective Transformer (DTST) tailored for AGPReID, which dynamically selects pivotal tokens to concentrate on pertinent regions. Specifically, we segment the input image into multiple tokens, with each token representing a unique region or feature within the image. Using a Top-k strategy, we extract the k most significant tokens that contain vital information essential for identity recognition. Subsequently, an attention mechanism is employed to discern interrelations among diverse tokens, thereby enhancing the representation of identity features. Extensive experiments on benchmark datasets showcases the superiority of our method over existing works. Notably, on the CARGO dataset, our proposed method gains 1.18% mAP improvements when compared to the second place. In addition, we comprehensively analyze the impact of different numbers of tokens, token insertion positions, and numbers of heads on model performance.**

***Index Terms*—Aerial Ground Person Re-identification, Top-k Token Selective Transformer, Attention Mechanism**

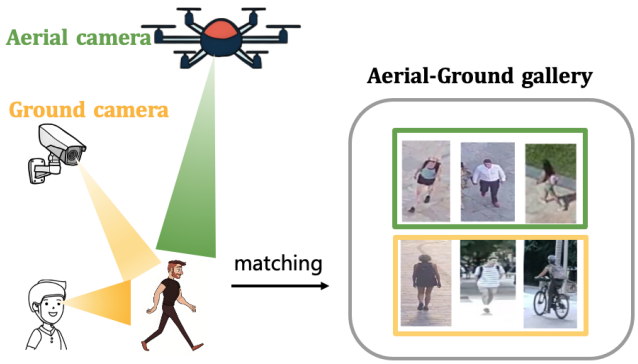

Fig. 1: A straightforward description of Aerial-Ground Person Re-identification (AGPReID) involves the utilization of an aerial-ground mixed camera network, enabling matching across aerial-aerial, ground-ground, and aerial-ground scenarios. Thus, it presents greater challenges and practical applications compared to traditional single-camera person ReID methods.

I. Introduction

Person Re-identification (ReID) is crucial for surveillance and tracking, identifying individuals across camera views. Advances in deep learning have improved feature extraction and matching accuracy [1]–[5]. However, most methods rely on global image features, making them vulnerable to background noise and irrelevant regions, particularly in cases of occlusion or complex backgrounds. This limits their effectiveness in diverse real-world scenarios with cross-camera variations and environmental inconsistencies [6]–[8].

To address these challenges, recent studies have emphasized the importance of more targeted and efficient feature extraction approaches. For instance, Zhang et al. [9] propose a separable attention mechanism to focus on discriminative regions while suppressing irrelevant background features. Tang et al. [10] introduce adaptive context-aware selection to dynamically enhance feature representations under complex conditions. Similarly, Qiu et al. [11] develop a salient feature extraction framework that prioritizes key object parts even in scenarios involving significant occlusion. These advancements show promising progress in overcoming the limitations of the reliance on global feature in View-homogeneous person ReID. However, when applied to Aerial-Ground Person Re-identification (AGPReID) tasks (View-heterogeneous person ReID), which are valuable in real-world scenarios for addressing complex aerial-to-ground matching challenges and encompassing diverse camera perspectives [12], these methods often fall short. Fig. 1 demonstrates the AGPReID problem. This discrepancy may stem from the scale diversity and redundancy characteristics observed in large-area observational scenarios, leading to notable appearance differences for the same individual across various cameras. Therefore, there is an urgent need to develop innovative strategies that effectively address these specific challenges in AGPReID.

To this end, we propose a Dynamic Token Selective Transformer (DTST) that enhances identity representation by focusing on the most critical spatial features. Our DTST module contains two steps: First, a Predictor Local-Global network computes relevance scores for each token, integrating local and global spatial semantics using multi-head attention. Second, a Perturbation-Based Top-K Selector chooses the most relevant tokens based on the predicted scores, ensuring robustness by adding noise perturbations. The selected tokens are combined with a global class token, enabling efficient and compact rep-

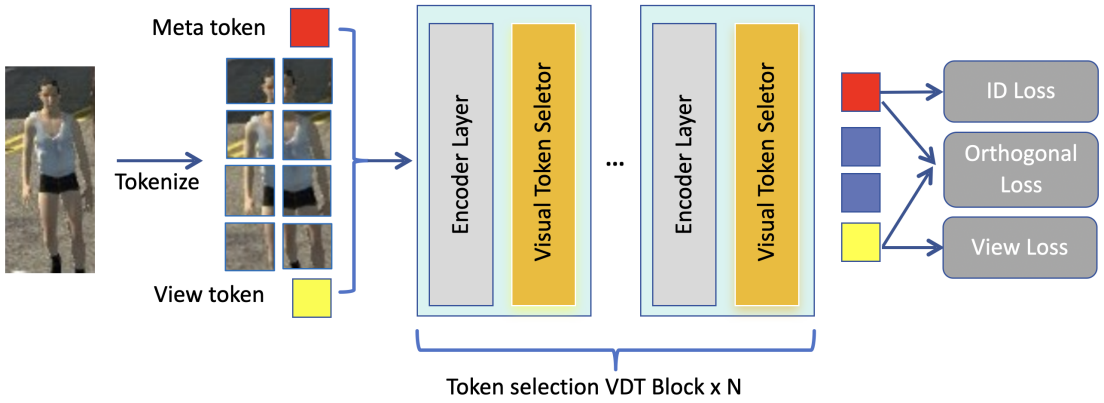

Fig. 2: Illustration of the proposed Dynamic Token Selective Transformer (DTST) framework. The framework incorporates N Token Selection view-decoupled transformer (VDT) blocks, where each block consists of an encoder layer and a visual token selector. The loss function is designed to account for both view-related and view-unrelated features, while an orthogonal loss ensures that these features remain independent from each other, further enhancing feature disentanglement and robustness.

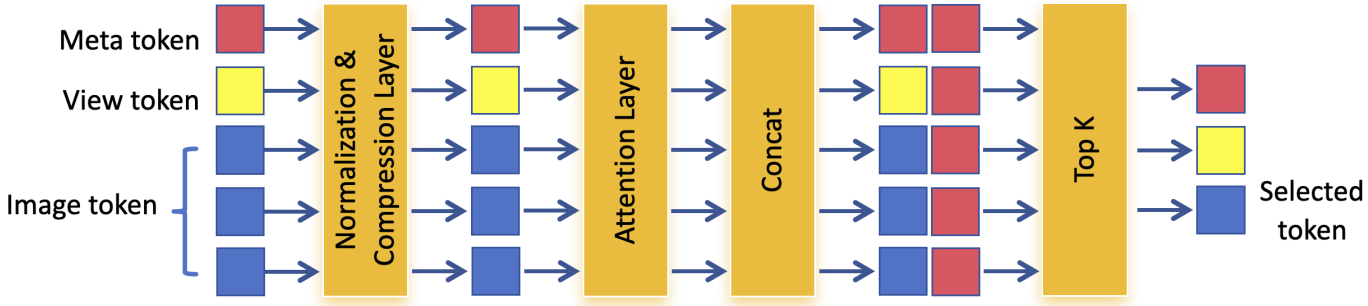

Fig. 3: The Illustration of Visual Token Selector (VTS). The process involves selecting the Top-K informative tokens from the original token set to be used in the subsequent feature aggregation.

resentation while reducing computational overhead. Extensive experiments validate our method's state-of-the-art performance on AGPReID tasks, showcasing its robustness in handling occlusions, complex backgrounds, and viewpoint variations.

Our main contributions are as follows.

- We propose a Top-k Token Selective Transformer for AGPReID, to better model identity representation spatially. We further comprehensively study the impact of the insertion position and the number of tokens selected on the model's performance.
- To eliminate the interference of irrelevant tokens, our method adaptively selects the most critical tokens based on the top-k selective mechanism, making the long-range modeling more effective and compact.
- Extensive experiments on various datasets demonstrate that our proposed model achieves state-of-the-art performance on AGPReID tasks.

II. Related Work

A. Person Re-identification

Person re-identification (ReID) is essential for retrieving images of the same individual across different camera views. It can be categorized into view-homogeneous and view-heterogeneous ReID. View-homogeneous ReID pertains to scenarios with a single camera type, such as ground-only or aerial-only networks, while view-heterogeneous ReID such as Aerial-Ground Person ReID (AGPReID), deals with networks featuring diverse camera perspectives. In terms of view-homogeneous ReID, ground-only camera networks have received more attention compared to aerial-only networks. For example, some ground-only datasets are well established such as Market1501 [13] and MSMT17 [14]. As a consequence, a multitude of methods have been proposed, such as hand-crafted feature-based, CNN-based, and transformer-based approaches, facilitating the development of ReID. However, these methods overlook the significant view differences between aerial and ground cameras, leading to poor performance faced with diverse view-point scenarios. Fortunately, view-heterogeneous ReID can address this issue. Recently, researchers in [12] propose the AG-ReID dataset, which includes identity and attribute labels, and put forward an attribute-guided model. Another work extends this by introducing the CARGO dataset with multiple matching scenarios and proposes a view-decoupled transformer (VDT) that decouples view-related features using hierarchical separation and orthog-

onal loss, improving performance and reducing reliance on extensive attribute labeling [15]. However, this approach does not dynamically select key tokens related to the target object, fails to reduce redundant computation, and lacks enhanced model capability to focus specifically on critical regions of interest.

B. Token Selection in Vision Transformers

Token selection is crucial for addressing redundancy issues in transformer-based vision models, particularly in tasks involving dense visual data. Despite their success, transformers often suffer from computational inefficiencies due to the need to process numerous redundant tokens. Token selection methods can effectively mitigate this issue by focusing on only the most informative tokens for further processing. For example, STTS [16], as a representative work, utilizes token selection to enhance computational efficiency by dynamically reducing the number of tokens processed at each transformer layer. These approaches have demonstrated substantial reductions in computation while maintaining performance. To address the challenge of differentiability in token selection, a perturbed maximum strategy is introduced [17], enabling top-K selection to be differentiable, thereby facilitating end-to-end training. Building on the principles of differentiable top-K selection [18], we develop a lightweight token selection module specifically designed to enhance temporal-spatial modeling in our view-decoupled transformer. By selecting only the most informative tokens, this module reduces redundancy and improves both efficiency and performance, especially in modeling visual data across multiple viewpoints.

III. Method

A. Formulation

Aerial-Ground Person ReID aims to match images from ground- or aerial-only camera networks. In a training dataset $\mathcal{D}^{tr} = \{(x_i, y_i, v_i)\}_{i=1}^{|\mathcal{D}^{tr}|}$, each instance consists of an image x_i depicting a person, along with identity label y_i and view label v_i. The view label $v_i \in \{v^a, v^g\}$ is determined by the known camera labels in $\mathcal{D}$, distinguishing between aerial (v^a) and ground (v^g) views. A substantial distinction in views between v^a and v^g results in a biased feature space, characterized by low intra-identity similarity and high inter-identity dissimilarity.

B. Overview

As illustrated in Fig.2, we propose a token enhanced framework based on the View-Decoupled Transformer (VDT) to tackle the view discrepancy challenge in AGPReID. Input images that include both aerial (v_a) and ground (v_g) views are tokenized into a sequence of tokens. To encompass both global and view-specific details, meta tokens and view tokens are added to these image tokens before they are inputted into our VDT.

Comprising N blocks, the VDT framework initiates each block with a conventional self-attention encoding process, succeeded by a subtraction operation between meta and view tokens to explicitly disentangle view-specific characteristics from the overarching ones. This facilitates a distinct segregation of features influenced by diverse viewpoints.

Subsequently, the updated meta and view tokens produced by the VDT are supervised by identity and view classifiers. To enforce the independence of meta and view tokens, we introduce an orthogonal loss, facilitating the successful separation of view-based and view-agnostic attributes. To select the most critical tokens, a **visual token selector** module is proposed to enhance the identity representation, with further elaboration provided in subsequent sections.

We introduce the Visual Token Selector (VTS), as shown in Fig. 3, designed to dynamically refine the token representation by selecting the most informative tokens for subsequent analysis. This module aims to reduce redundancy and enhance the model's ability to focus on critical regions, thereby optimizing computational efficiency while preserving feature quality. The VTS mechanism can be understood as a dynamic token selection process that leverages attention scores to determine the importance of each token.

For a sequence of tokens $\{t_i\}_{i=1}^{M}$, where M is the number of tokens, the VTS computes importance scores for each token s_i using a lightweight attention mechanism. The score s_i is obtained as:

$$s_i = \text{softmax}\left(\frac{t_i^\top W_q W_k^\top t_i}{\sqrt{d}}\right),$$

where t_i is the i-th token, W_q and W_k are learnable matrices representing query and key transformations, and d is the dimensionality of the tokens. The softmax function normalizes the scores to ensure they sum to 1, thus creating a probabilistic distribution over the tokens.

These tokens are then ranked based on their importance scores, and we select the top-K tokens with the highest scores, where $K < M$ is a hyperparameter that controls the number of tokens retained. Mathematically, this selection can be represented as:

$$\{t_i^{\text{selected}}\} = \text{TopK}(\{s_i\}_{i=1}^{M}),$$

where $\text{TopK}(\cdot)$ returns the indices corresponding to the top-K scores. The retained tokens, $\{t_i^{\text{selected}}\}$, are then passed to the subsequent layers or directly to the final classification head.

To ensure that the VTS can be used in an end-to-end training fashion, we adopt a differentiable approach for the token selection. Specifically, we use a continuous relaxation of the TopK function by employing a Gumbel-Softmax trick:

$$\hat{s}_i = \frac{\exp((s_i + g_i)/\tau)}{\sum_{j=1}^{M} \exp((s_j + g_j)/\tau)},$$

where g_i are Gumbel noise samples and τ is the temperature parameter that controls the smoothness of the approximation. This differentiable approximation allows the selection of tokens to be included in backpropagation, facilitating end-to-end optimization.

By incorporating the Visual Token Selector, we achieve several key benefits:

- **Reduce redundancy**: By selecting only the most informative tokens, we minimize the amount of redundant information processed by the model.
- **Enhance discriminability**: The model can focus on the most critical aspects of the input, leading to improved performance on tasks requiring fine-grained feature analysis.
- **Improve computational efficiency**: Reducing the number of tokens processed helps in reducing the overall computational cost, making the model more efficient for both training and inference.

IV. Experiments

A. Experiment settings

Datasets. We conduct experiments on the CARGO [15] dataset and AG-ReID dataset [19]. Compared to AG-ReID, the CARGO dataset offers a larger scale, greater diversity, and is the first large-scale synthetic dataset for AGPReID. Table I summarizes both datasets. For CARGO, 51,451 images with 2,500 IDs are used for training, and 51,024 images with 2,500 IDs for testing. Four evaluation protocols (ALL, A↔A, G↔G, and A↔G) assess model performance, with A↔A and G↔G testing aerial and ground data separately, and A↔G using cross-view retrieval. The training set is consistent across all protocols.

For AG-ReID, 11,554 images with 199 IDs are used for training, and 12,464 images with 189 IDs for testing. Two protocols, A→G and G→A, evaluate cross-view retrieval, with the former testing 1,701 aerial queries against 3,331 ground galleries, and the latter 962 ground queries against 7,204 aerial galleries.

Evaluation Metrics. Following the common setting, we utilize three metrics to evaluate our model: the cumulative matching characteristic at Rank1, mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP).

B. Implementation Details

Our model is implemented using the PyTorch framework, with experiments conducted on an NVIDIA 4090 GPU. We use the View-decoupled Transformer (VDT) as the baseline, which includes 12 transformer encoder blocks based on the ViT-Base architecture, pre-trained on ImageNet with a patch size and stride of 16×16. Input images are resized to 256×128 during preprocessing. The training process employs the Stochastic Gradient Descent (SGD) optimizer with a cosine learning rate decay, starting at 8×10^{-3} and reducing to 1.6×10^{-6} over 120 epochs. The batch size is set to 128, comprising 32 identities with four images per identity. Our token selector module features a two-head transformer encoder that selects the top two rated tokens, integrated after the final transformer encoder block for enhanced performance.

C. Comparisons with State-of-the-art Methods

We evaluate our proposed DTST against state-of-the-art methods on the CARGO and AG-ReID datasets, comprising CNN-based approaches (BoT [21], SBS [20], MGN [22], AGW [23]) and transformer-based methods (ViT [24], VDT [15]).

Performance on CARGO. Table II shows the results of our proposed DTST and other competitive methods on the CARGO dataset. The proposed DTST achieves state-of-the-art performance. For example, DTST surpasses the mAP/Rank-1/mINP of the baseline by 1.18%/3.13%/0.43% on the aerial-to-ground (A↔G) protocol of CARGO. Besides, DTST also brings different degree of benefits to other CARGO protocols. Specifically, our proposed DTST exceeds VDT on mAP/Rank-1/mINP by 1.51%/1.60%/2.00% on the ALL of AG-ReID. Demonstrating the effectiveness of the Dynamic Token Selective Transformer in mitigating view bias and improving identity representation. Previous view-homogeneous ReID methods show significant performance degradation under the view-heterogeneous AGPReID protocols, especially in cases of considerable view variation. This decline underscores how view bias hampers the consistency of identity features across views. Unlike existing methods that overlook this key challenge and struggle to generalize in heterogeneous scenarios, our approach adaptively selects the most critical tokens using a top-k selective mechanism. This token selection not only maintains accuracy but even enhances it, resulting in more effective and compact long-range modeling.

Performance on AG-ReID. To further demonstrate the performance of our model, we also carry out similar experiments on the AG-ReID dataset. The outcomes are detailed in Table III. As depicted in Table III, we compare two challenging protocols: A→G and G→A. It is noteworthy that VDT serves as a strong baseline. However, our proposed method, DTST, demonstrates a significant enhancement, outperforming VDT by 0.57% for the A→G Rank-1 protocol and 1.04% for the G→A Rank-1 protocol. This consistent improvement suggests that the superior performance of DTST does not stem from a robust baseline VDT but from the proposed method itself.

D. Ablation Study

In this section, we provide ablation study to investigate several key components of our DTST. We also delved into the number of attention heads, token quantities, and token positions. Notably, all ablation experiments are conducted on the on the CARGO dataset.

Effects of Visual Token Selector (VTS). We first explore the effectiveness with placing the Visual Token Selector before the final layer of the View-decoupled Transformer. In this setup, all other settings,such as the number of attention heads and selected tokens, remain constant. Table IV shows the results, where model-a lacks a visual token selector, whereas model-b incorporates one. From the Table, we can observe a 5.63% improvement in rank-1 accuracy and 1.34% increase in mAP accuracy under Protocol A→G, which indicates that the token

TABLE I: THE DETAILED SUMMARY OF THE DATASET PROPERTIES INVOLVED IN THIS PAPER, INCLUDING AG-ReID and CARGO.

Dataset	Data	#PersonID	#Camera	#Image	#Height
AG-ReID [19]	Real	388	2 (1A+1G)	21,893	$15 \sim 45m$
CARGO [15]	Synthetic	5,000	13 (5A+8G)	108,563	$5 \sim 75m$

TABLE II: Performance comparison of the mainstream methods under four settings of the proposed CARGO dataset. "ALL" denotes the overall retrieval performance of each method. "G↔G," "A↔A," and "A↔G" represent the performance of each model in several specific retrieval patterns. Rank1, mAP, and mINP are reported (%). The best performance is shown in **bold**.

Method	Protocol 1: ALL			Protocol 2: G↔G			Protocol 3: A↔A			Protocol 4: A↔G		
	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP
SBS [20]	50.32	43.09	29.76	72.31	62.99	48.24	67.50	49.73	29.32	31.25	29.00	18.71
PCB [37]	51.00	44.50	32.20	74.10	67.60	55.10	55.00	44.60	27.00	34.40	30.40	20.10
BoT [21]	54.81	46.49	32.40	77.68	66.47	51.34	65.00	49.79	29.82	36.25	32.56	21.46
MGN [22]	54.81	49.08	36.52	**83.93**	71.05	55.20	65.00	52.96	36.78	31.87	33.47	24.64
VV [40, 41]	45.83	38.84	39.57	72.31	62.99	48.24	67.50	49.73	29.32	31.25	29.00	18.71
AGW [23]	60.26	53.44	40.22	81.25	71.66	58.09	67.50	56.48	40.40	43.57	40.90	29.39
ViT [24]	61.54	53.54	39.62	82.14	71.34	57.55	80.00	**64.47**	**47.07**	43.13	40.11	28.20
VDT [15]	62.82	54.22	39.92	79.46	70.60	57.89	**82.50**	64.06	44.67	47.50	42.21	29.03
DTST (Ours)	**64.42**	**55.73**	**41.92**	78.57	**72.40**	**62.10**	80.00	63.31	44.67	**50.63**	**43.39**	**29.46**

TABLE III: Quantitative evaluation of the mainstream methods under two settings of AG-ReID dataset. "A↔G", and "G↔A" represent the performance in two specific patterns. Rank1, mAP, and mINP are reported (%). Best marked in **bold**.

Method	Protocol 1: A→G			Protocol 2: G→A		
	Rank1	mAP	mINP	Rank1	mAP	mINP
SBS [20]	73.54	59.77	-	73.70	62.27	-
BoT [21]	70.01	55.47	-	71.20	58.83	-
OSNet [25]	72.59	58.32	-	74.22	60.99	-
ViT [24]	81.28	72.38	-	82.64	73.35	-
VDT [15]	82.91	74.44	**51.06**	83.68	75.96	49.39
DTST (ours)	**83.48**	**74.51**	49.86	**84.72**	**76.05**	**50.04**

TABLE IV: Ablation study of model key designs on CARGO dataset. Rank1, mMAP, and mINP are reported(%). Best in **blod**.

Method	Visual Token Selector	Protocol: A↔G		
		Rank1	mAP	mINP
model-a	✘	45.00	42.05	**30.26**
model-b (**Ours**)	✔	**50.63**	**43.39**	29.46

selection strategy effectively filters out tokens with discriminative features and eliminates identity-irrelevant tokens, thereby enhancing better identity representation.

Number of Heads. We also evaluate the performance of VTS with different numbers of heads, specifically 2, 4, and 8 heads in Table V. Interestingly, using more heads results in a decrease in accuracy. Specifically, when increasing the number of heads from 2 to 4, there is a 3.64% decline in rank-1 accuracy and 0.93% drop in mAP. This suggests that a higher number of heads may dilute the model's ability to focus on critical identity features, potentially introducing noise and decreasing overall model performance. One underlying reason may be model over-fitting, as a greater number of heads could increase the model's complexity without corresponding improvements in performance. Another potential explanation might be that more heads may dilute the importance of the most vital tokens, leading to less effective feature aggregation.

TABLE V: Ablation study on the number of attention heads, token quantities, and token positions using the CARGO dataset. "Head-Num." signifies the quantity of attention heads, "T-Num." demotes the number of token, and "T-Position." indicates the specific position where each token is locate. Performance is assessed through Rank1, mAP, and mINP(%), with the best results highlighted in **blod**.

Method	Head-Num.	T-Num.	T-Position.	Protocol: A↔G		
				Rank1	mAP	mINP
model-1	8	2	last layer	46.25	42.56	**30.16**
model-2	8	3	last layer	45.00	41.28	28.83
model-3	8	3	second to last layer	46.88	41.04	28.12
model-4	8	32	second to last layer	40.00	36.58	24.73
model-5	4	2	last layer	46.88	42.46	29.79
model-6 (**Ours**)	2	2	last layer	**50.63**	**43.39**	29.46

Number of Tokens Selected. Keeping other variables constant, we analyze the impact of different numbers of token selections on model performance in Table V. We vary the number of tokens to 2, 3, 5. The findings reveal that selecting 2 or 3 tokens yields superior results across all evaluation metrics, i.e. Rank-1 accuracy, mAP, and mINP. Specifically, we increase the number of selected tokens beyond 3, but the performance fails to show any improvement, indicating that opting for fewer but more critical tokens enables the model to concentrate better on pivotal identity features. In contrast, selecting more tokens may introduce irrelevant information, thereby compromising overall accuracy. When our method is applied in the same setup, choosing 3 tokens compared to 2 tokens results in a decrease of 1.25% in rank-1 accuracy,

1.28% in mAP, and 1.33% in mINP, highlighting the trade-off between token quantity and model's focus on essential features.

Token positions. The insertion position of VST, whether in the last or second-to-last layer, also affects model performance, as shown in Table V. When the fixed number of heads is 8 and the number of tokens is 3, model-3 achieves a higher Rank-1 accuracy at 46.88%, but both mAP and mINP decrease. The reason behind this could be that tokens in shallow layers contain more detailed information, while tokens in deeper layers extract higher-level semantic information. As a result, the information within each token becomes more refined, leading to a higher compressibility ratio.

V. Conclusion and Future Work

In this paper, we investigate the relationships between tokens in transformers and propose a dynamic token selective transformer specifically for the AGReID task. Experiments demonstrate that incorporating token selection can effectively reduce token redundancy, enhance the importance of discriminative tokens, and consequently achieve state-of-the-art results. Furthermore, we investigated the impact of different implementation details, the number of tokens, and the position of token insertion on model performance, providing a comprehensive understanding of the influence of token selection on AGReID. Token selection is a general technique, and we will explore its application in other tasks. While our work focuses on token-level selection, recent studies demonstrate the potential of pixel-level operations [26], showing effectiveness in tasks like object classification, masked autoencoding, and image generation. Inspired by this, we aim to integrate token and pixel selection to enhance the efficiency and performance of vision models.

References

[1] Xiaoxiao Sun and Liang Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 608–617.

[2] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie, "Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4288–4298.

[3] Jianxiong Tang, Jian-Huang Lai, Xiaohua Xie, and Lingxiao Yang, "Spike count maximization for neuromorphic vision recognition.," in *IJCAI*, 2023, pp. 4253–4261.

[4] Pengze Zhang, Lingxiao Yang, Xiaohua Xie, and Jianhuang Lai, "Pose guided person image generation via dual-task correlation and affinity learning," *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[5] Qi Chen, Yun Chen, Yuheng Huang, Xiaohua Xie, and Lingxiao Yang, "Region-based online selective examination for weakly supervised semantic segmentation," *Information Fusion*, vol. 107, pp. 102311, 2024.

[6] Geon Lee, Sanghoon Lee, Dohyung Kim, Younghoon Shin, Yongsang Yoon, and Bumsub Ham, "Camera-driven representation learning for unsupervised domain adaptive person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11453–11462.

[7] He Li, Mang Ye, and Bo Du, "Weperson: Learning a generalized re-identification model from all-weather virtual data," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 3115–3123.

[8] Quan Zhang, Jian-Huang Lai, Zhanxiang Feng, and Xiaohua Xie, "Uncertainty modeling with second-order transformer for group re-identification," in *Proceedings of the AAAI conference on artificial intelligence*, 2022, vol. 36, pp. 3318–3325.

[9] Quan Zhang, Jianhuang Lai, Xiaohua Xie, Xiaofeng Jin, and Sien Huang, "Separable spatial-temporal residual graph for cloth-changing group re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[10] Jianxiong Tang, Jian-Huang Lai, Xiaohua Xie, Lingxiao Yang, and Wei-Shi Zheng, "Ac2as: Activation consistency coupled ann-snn framework for fast and memory-efficient snn training," *Pattern Recognition*, vol. 144, pp. 109826, 2023.

[11] Junyang Qiu, Zhanxiang Feng, Lei Wang, and Jianhuang Lai, "Salient part-aligned and keypoint disentangling transformer for person re-identification in aerial imagery," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[12] Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes, "Aerial-ground person re-id," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 2585–2590.

[13] 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*, 2015, pp. 1116–1124.

[14] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 79–88.

[15] Quan Zhang, Lei Wang, Vishal M Patel, Xiaohua Xie, and Jianhaung Lai, "View-decoupled transformer for person re-identification under aerial-ground camera network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22000–22009.

[16] Junke Wang, Xitong Yang, Hengduo Li, Li Liu, Zuxuan Wu, and Yu-Gang Jiang, "Efficient video transformers with spatial-temporal token selection," in *European Conference on Computer Vision*. Springer, 2022, pp. 69–86.

[17] Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach, "Learning with differentiable pertubed optimizers," *Advances in neural information processing systems*, vol. 33, pp. 9508–9519, 2020.

[18] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin, "Ts2-net: Token shift and selection transformer for text-video retrieval," in *European conference on computer vision*. Springer, 2022, pp. 319–335.

[19] Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes, "Agreid.v2: Bridging aerial and ground views for person re-identification," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 2896–2908, 2024.

[20] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei, "Fastreid: A pytorch toolbox for general instance re-identification," in *ACM Int. Conf. Multimedia*, 2023, pp. 9664–9667.

[21] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[22] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 274–282.

[23] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, 2022.

[24] Alexey Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[25] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang, "Learning generalisable omni-scale representations for person re-identification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5056–5069, 2021.

[26] Duy-Kien Nguyen, Mahmoud Assran, Unnat Jain, Martin R Oswald, Cees GM Snoek, and Xinlei Chen, "An image is worth more than 16x16 patches: Exploring transformers on individual pixels," *arXiv preprint arXiv:2406.09415*, 2024.