CAViT: Contextual Alignment Vision Transformer for Video Object Re-identification

Jinlin Wu[1,2,3], Lingxiao He[5], Wu Liu[4], Yang Yang[1,2], Zhen Lei[1,2,3]*, Tao Mei[4], and Stan Z. Li[6]

[1] CBSR & NLPR, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] Centre for Artificial Intelligence and Robotics, HKISI, CAS
[4] JD Explore Academy
[5] Longfor Inc.
[6] School of Engineering, Westlake University, Hangzhou, China
{jinlin.wu, yang.yang, zlei}@nlpr.ia.ac.cn, {liuwu1, tmei}@jd.com, xiaomingzhidao1@gmail.com, Stan.ZQ.Li@westlake.edu.cn

Abstract. Video object re-identification (reID) aims at re-identifying the same object under non-overlapping cameras by matching the video tracklets with cropped video frames. The key point is how to make full use of spatio-temporal interactions to extract more accurate representation. However, there are dilemmas within existing approaches: (1) 3D solutions model the spatio-temporal interaction but are often troubled with the misalignment of adjacent frames, and (2) 2D solutions adopt a divide-and-conquer strategy against the misalignment but cannot take advantage of the spatio-temporal interactions. To address the above problems, we propose a Contextual Alignment Vision Transformer (**CAViT**) to the spatio-temporal interaction with a 2D solution. It contains a Multi-shape Patch Embedding (**MPE**) module and a Temporal Shift Attention (**TSA**) module. MPE is designed to retain spatial semantic information against the misalignment caused by pose, occlusion, or misdetection. TSA is designed to achieve contextual spatial semantic feature alignment and jointly model spatio-temporal clues. We further propose a Residual Position Embedding (**RPE**) to guide TSA in focusing on the temporal saliency clues. Experimental results on five video person reID datasets demonstrate the superiority of the proposed CAViT. Additionally, the experiment conducted on VVeRI-901-trial also shows the effectiveness of CAViT for the video vehicle reID. Our code is available on https://github.com/KimWu1994/CAViT.

Keywords: Video object reID, Vision transformer, Temporal shift attention, Residual position embedding

1 Introduction

Video object re-identification (reID) is a challenging task which matches video tracks of objects across non-overlapping cameras. The spatio-temporal re-

* Corresponding author.

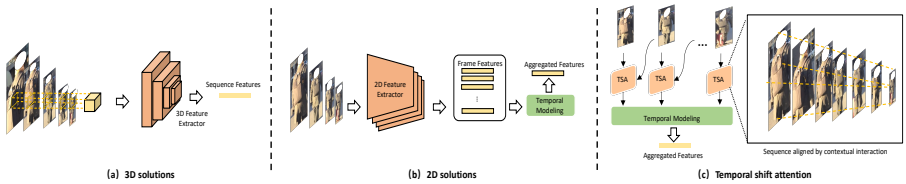

Fig. 1: Illustration of various solutions applied in spatio-temporal learning. (a) 3D solutions. (b) 2D solutions. (c) The proposed temporal shift attention jointly models spatio-temporal clues against the misalignment of adjacent frames.

lation information of video tracklets often contains diverse viewpoints and pose variations. Thus, how to learn accurate and robust spatio-temporal representations in a video track is a crucial component for video object reID.

Many existing methods as shown in Fig. 1(a) apply 3D convolutional neural networks to learn spatio-temporal features in a sequence of video frames. Although it can integrate feature extraction and temporal modeling in one step, it is inevitably affected with spatial misalignment caused by the movement of objects. To this end, some 2D solutions in Fig. 1(b) attempt to adopt a divide-and-conquer strategy that tackles feature representation and feature aggregation separately. However, the divide-and-conquer strategy cannot take full advantage of spatio-temporal interactions.

In this paper, we propose **C**ontextual **A**lignment **V**ision **T**ransformer (**CAViT**) which learns accurate and robust spatial-temporal features. Firstly, we replace the self-attention of ViT [9] with a Temporal-Shift Attention (TSA) to align the objects of adjacent frames. It naturally transfers the spatio-temporal modeling task from a 3D representation learning problem to a 2D contextual alignment problem, as shown in Fig. 1(c). To further guide TSA in focusing on the temporal saliency region, we propose a novel yet effective residual position embedding module (RPE) which utilizes the relative variation of the adjacent frames denoting the temporal position. We also design a multi-shape patch embedding (MPE) that provides rich semantic information to improve the ability of feature representation. Experiments on video person reID and vehicle reID show that CAViT achieves relatively high performance even in the presence of heavy occlusion and misdetection. Moreover, CAViT significantly outperforms the state-of-the-arts on video person / vehicle reID benchmarks. Especially, on LSVID and PRID2011, CAViT respectively achieves 89.3% rank1 and 97.5% rank1 performance.

Generally speaking, the main contributions of this paper are as follows:

- We propose a novel video representation learning framework CAViT for video object reID, which jointly learns accurate and robust spatio-temporal features with a 2D vision transformer model.
- We propose a new temporal shift attention module to replace the self-attention mechanism of the vision transformer. It aligns the adjacent frames to extract accurate pedestrian representations from an entire sequence.

– We develop a multi-shape patch embedding module to improve the scalability of the vision transformer. A novel residual position embedding is also introduced to guide our model in focusing the temporal saliency information among consecutive frames.

2 Related work

2.1 Video Re-identification

The research of video reID has made great progress. As shown in Fig. 2, the rank1 accuracy improves from 30.7% to 91.5% in recent years on MARS dataset. We mainly review highly related video-reID methods in this subsection and give comparisons in Sec. 4 to show the superiority of our method on multiple datasets.

3D solutions. Some approaches consider video reID as a spatio-temporal representation learning task. To make full use of the temporal clues, some 3D solutions (*e.g.*, C3D [41], P3D [40], SlowFast [11], I3D [44]) are introduced to video reID. However, due to the misalignment of adjacent frames, 3D CNNs are troubled with the background and occlusion. In order to solve this, some 3D alignment convolutional layers are proposed. For example, Li *et al.* [26] design a two-branch 3D CNN network, where one branch is used to capture optical flow clues and the other is used for spatial clues. Another approach is to develop a 3D non-local module (*e.g.*, AP3D [13], Bicnet-tks [19], RFCne [22]) and insert this module to 3D CNNs for the alignment of adjacent frames. However, limited by the locality of the convolution, these methods only align the local region of adjacent frames and cannot solve the misalignment of the whole frame.

2D solutions. Other approaches treat video reID as a set representation learning task. To obtain the set representation, some divide-and-conquer based strategies are proposed. They firstly apply 2D CNNs as the feature extractor to obtain features of each frame and then use a feature post process module (*e.g.*, average/maximum temporal pooling, recurrent neural networks (RNNs), or attention mechanisms) to obtain the set average feature. Zhou *et al.* [54]

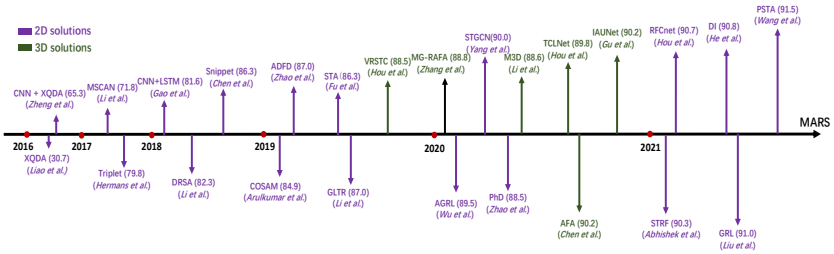

Fig. 2: Development of video reID methods on MARS. The number in parentheses for each method represents the corresponding rank1 performance.

apply a RNN to aggregate multiple frame features. Yang *et al.* [47] propose a spatial-temporal graph convolution network to model the temporal relations of different frames and spatial relations within a frame. Abandoning temporal interaction can free models from the misalignment of adjacent frames [1] but it may disregard some important temporal clues. So PSTA [43] uses aggregation module for ID switch problem. As discussed in above, all those methods cannot tackle temporal dependency, attention, and spatial misalignment simultaneously.

2.2 Transformer based reID

Transformer breaks the locality limitation of the convolution model, and shows its superiority over convolutional architectures in many vision tasks like image classification and object detection, *e.g.*, DETR [4]. These methods are designed based on the encoder-decoder architecture of the transformer, which applies queries to read the target information from the encoding representations. However, the decoder may be not the necessary component for the visual representation learning task. The decoder-free methods are then proposed, named vision transformer, *e.g.*, ViT [9], Cross ViT [6] and Swin transformer [36]. These methods mainly adopt a patch embedding module and a self-attention mechanism for visual representation learning.

Benefiting from the development of the transformer, the object reID task also makes great progress. For the image-based object reID task, Li *et al.* [30] introduce the vanilla transformer into the partial person reID task, in which the decoder applies K queries for robust representations against the misalignment caused by occluded and partial situations. Liao *et al.* [31] propose a pair-based cross-attention strategy. They use the transformer decoder as a feature post-processing module to re-fine the similarity score of the probe-gallery pair in the unseen scene. Several vision transformer-based methods are also applied to the image-based reID task. He *et al.* [16] propose a ViT based object reID model. To learn representations suitable for cross-camera retrieval, it proposes several strategies including camera position embedding, overlapping patch embedding, jigsaw patch module, etc. Zhu *et al.* [55] propose an auto-aligned strategy in vision transformer to alleviate the misalignment of the feature matching. For the video reID task, He *et al.* [17] design a dense interaction method for transformer to obtain robust embedding. However, it is difficult for training, *i.e.,* dense interaction needs 4 GPUs and 800 epochs for convergence.

3 Methodology

3.1 Problem Formulation

Video object reID aims to retrieve the same object with a query sequence from a gallery set. Let denote $\mathcal{P}$ as the query sequence and $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_K\}$ as the gallery set, where it contains K sequences and each sequence has multiple images. Corresponding features $f_{\mathcal{G}_k}$ for a gallery sequence and $f_{\mathcal{Q}}$ for the query

sequence can be extracted by a video feature learning network. Video object reID retrieves the target gallery video $\mathcal{G}$ that is the most similar to the query in the video representation space, *i.e,*

$$\mathcal{G} = \max_{\mathcal{K}} \mathcal{S}(f_{\mathcal{G}_k}, f_{\mathcal{Q}}), \tag{1}$$

where $\mathcal{S}$ is the similarity score of the gallery and query sequences. The key of this task is how to extract discriminative representations from the given sequence $\mathcal{T} = \{I^1, \cdots, I^{\mathcal{N}}\}$.

$$f^{\mathcal{T}} = \phi(I^1, \cdots, I^{\mathcal{N}}), \tag{2}$$

where ϕ is a model extracting discriminative representation from spatio-temporal clues of the video.

3.2 Contextual Alignment Feature Learning

There are two existing frameworks for designing ϕ: 3D solutions and 2D solutions. 3D solutions often apply the 3D CNN as the backbone to jointly learn representations from the whole sequence, as follow:

$$f^{\mathcal{T}} = \phi_{3D}(I^1, \cdots, I^{\mathcal{N}}), \tag{3}$$

where ϕ_{3D} denotes 3D CNN backbones. However, ϕ_{3D} is affected by the misalignment of adjacent frames and fails to extract precise representations.

To alleviate this problem, 2D solutions abandon contextual interaction in spatial clues modeling and adopt a divide-and-conquer strategy:

$$f^{\mathcal{T}} = \psi(\phi_{2D}(I^1), \cdots, \phi_{2D}(I^{\mathcal{N}})), \tag{4}$$

where ϕ_{2D} denotes 2D CNNs to extract representation for each frame. The 3D representation learning in Eq. 3 is divided into a spatial modeling module ϕ_{2D} and a temporal modeling module ψ. But the performance of ψ is limited, since there is no temporal interaction in ϕ_{2D}.

Considering the aforementioned dilemmas of 2D & 3D solutions, we model the sequence representation problem as a contextual alignment task, *i.e.,*

$$f((I^t|(I^1, \cdots, I^{\mathcal{N}})) = f(I^t|I^{t-1}). \tag{5}$$

Inspired by Markov chains [28], we focus on the dependencies between the current frame and the previous frame and propose an contextual alignment module $\mathcal{A}$. It models contextual interaction between x^t and x^{t-1}. The spatio-temporal joint modeling task can be formulated as follows:

$$\begin{aligned} f^{\mathcal{T}} &= \phi_{3D}(I^1, \cdots, I^{\mathcal{N}}) \\ &= \phi_{2D}(I^1) + \cdots + \phi_{2D}(\mathcal{A}(I^{\mathcal{N}}|I^{\mathcal{N}-1})) \\ &= \phi_{2D}(I^1) + \sum_{t=2}^{\mathcal{N}} \phi_{2D}(\mathcal{A}(I^t|I^{t-1})). \end{aligned} \tag{6}$$

According to Eq. 6, CAViT transfers the spatio-temporal joint modeling task to a contextual alignment problem. The 3D representation learning task of Eq. 3 can be reduced to a 2D representation learning task. The dilemmas between the contextual interaction modeling and the misalignment robustness are also be alleviated.

3.3 Contextual Alignment Vision Transformer

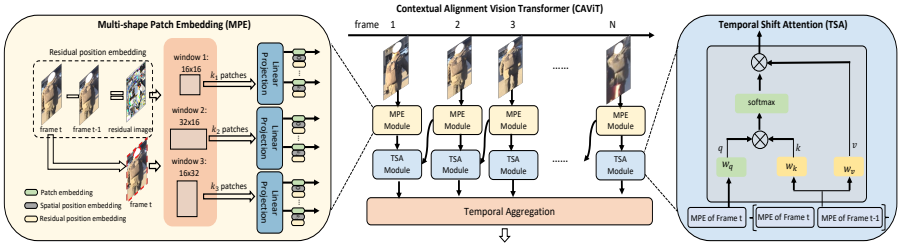

Fig. 3: Framework of **CAViT**. In the **multi-shape patch embedding module**, the input pedestrian sequence is divided into multi-shape patches with multi-shape windows and linear-projected to embeddings. The **residual position embedding** and the learnable 1D spatial position embedding are added to the patch embedding. The **temporal shift attention module** is applied to align the adjacent frames for joint spatio-temporal modeling.

Contextual Alignment Vision Transformer (CAViT) provides a feasible solution for spatio-temporal joint learning. An overview of the CAViT is presented in Fig. 3. The main pipeline of CAViT can be formulated as:

$$
\begin{aligned}
x_0^t &= \mathbf{MPE}(I^t) + \mathcal{R} + \mathcal{P} \\
\hat{x}_l^t &= [x_l^t; x_l^{t-1}] \\
y_l^t &= x_l^t + \mathbf{TSA}(\mathbf{LN}(x_l^t), \mathbf{LN}(\hat{x}_l^t)) \\
x_{l+1}^t &= y_l^t + \mathbf{FFN}(\mathbf{LN}(y_l^t)).
\end{aligned} \tag{7}
$$

Given a pedestrian sequence$\{I^1, \cdots, I^N\}$, the multi-shape patch embedding module MPE embeds the frame to multi-shape embedding vectors with different shaped windows. Then, a learnable 1D vector $\mathcal{P}$ and the residual position embedding $\mathcal{R}$ are added to the patch embeddings. The former position embedding denotes the spatial position in the current frame, while the latter indicates the temporal variation of the current frame. After these steps, we get the patch embedding x_0^t, which is the input of temporal shift attention layers TSA in CAViT. TSA is the alignment module $\mathcal{A}$ in Eq. 6 used to align current frame I^t and the previous frame I^{t-1}. The attention mechanism of CAViT is built by stacking

TSA. FFN and LN are the feed-forward network and Layer normalization of transformer attention block, respectively.

Residual Position Embedding (RPE). ViT is insensitive to the input order and treats each frame of the input sequence equally. Thus, attention power is wasted by redundant information in consecutive frames. To address this problem, we propose a residual position embedding to guide the model in focusing the temporal saliency information, as follows:

$$\mathcal{R}_{s_i}(I^t) = \mathcal{F}_{s_i}(\text{SoftMax}(I^t - I^{t-1})), \tag{8}$$

where $\mathcal{F}_{s_i}$ is the linear projection of the shape s_i. It encodes the residual of the i-th frame and the previous $(i-1)$-th frame as the position embedding. Softmax is used to normalize the residuals signal, suppress signals with small variations and amplify those with large variations caused by viewpoint changing, scale changing, and occlusions. Benefiting from MPE, CAViT extracts diversity information and learns robust representations.

Multi-shape Patch Embedding (MPE). The 16x16 patch is not scaleable enough in the origin ViT model. To perceive objects at different scales, we propose a multi-shape patch embedding module as follows:

$$x_0^t = \mathcal{F}_{s_i}(I^t) + \mathcal{P}_{s_i} + \mathcal{R}_{s_i}(I^t), \tag{9}$$

where $\mathcal{F}^{s_i}$ is the linear projection module of the i-th shape. $\mathcal{P}_{s_i}$ is the spatial position embedding. We adopt the learnable position embedding method as [9], allotting a learnable 1D vector P_{s_i} for each patch at the s_i shape. $\mathcal{R}_{s_i}(I^t)$ is the temporal position embedding as Eq. 8.

Temporal Shift Attention. For a sequence, p_i^t is the i-th patch of I^t, the t-th frame in the pedestrian sequence.

$$\begin{aligned} q_i^t &= p_i^t * W_q \\ k_i^t &= p_i^t * W_k \\ q_i^t &= p_i^t * W_v, \end{aligned} \tag{10}$$

where W_q, W_k, and W_v are the linear function. q_i^t, k_i^t and v_i^t are the inputs of the attention machine, respectively. The temporal shift attention (TSA) can be modeled as:

$$\begin{aligned} TSA(p_i^t) &= \text{Softmax}(q_i^t \times \mathcal{K}) \times \mathcal{V} \\ \mathcal{K} &= [k_1^t, \ldots, k_N^t, k_1^{t-1} \ldots, k_N^{t-1}]^T \\ \mathcal{V} &= [v_1^t, \ldots, v_N^t, v_1^{t-1}, \ldots, v_N^{t-1}]. \end{aligned} \tag{11}$$

Suppose the normalization factor of SoftMax is γ, TSA can be formulated as:

$$\begin{aligned} TSA(p_i^t) &= \frac{1}{\gamma}[q_i^t * {k_1^t}^T, \ldots, q_i^t * {k_N^{t-1}}^T)] \times [v_1^t, \ldots, v_N^{t-1}] \\ &= \frac{q_i^t * {k_1^t}^T}{\gamma} * v_1^t + \cdots + \frac{q_i^t * {k_N^{t-1}}^T}{\gamma} * v_N^{t-1} \\ &= \sum_{k=k_1^{t-1}, v=v_1^{t-1}}^{k=k_N^t, v=v_N^t} \frac{q_i^t * k^T}{\gamma} * v, \end{aligned} \tag{12}$$

where K^T concatenate all the patches of I^{t-1} and I^t. $q_i^t * k^T$ computes the similarity the patch p_i^t and all patches of adjacent frames. The similarity of patch p_i^t and I^t is the intra-frame self-attention, while the similarity of patch p_i^t and I^{t-1} is the inter-frame interaction. Specifically, if p_i^t belongs to an occluder which appears suddenly at I^t and cannot align to I^{t-1}, the response of the occluder p_i^t will be weakened. On the contrary, if p_i^t aligns to it's previous frame, the response will be enhanced. For this reason, TSA is more robust to the ID switch noise in the video object reID.

4 Experiments

4.1 Experiment Implement

Datasets. We evaluate the proposed method on five video person reID datasets and a video vehicle reID dataset, *i.e.,* MARS [53], MARS_DL [37], LSVID [25], PRID-2011 [18], iLIDS-VID [42] and VVeRI-901-trial [23]. The details of these datasets are summarized in Tab. 1. The bounding boxes are detected with DPM detector [12], and tracked using the GMMCP tracker [8]. The misalignment caused by the DPM detector and ID switch by the GMMCP tracker leads to confusion of video reID models. Liu *et al.* [37] clean MARS as MARS_DL. They re-detect the pedestrian bounding boxes with YOLOV4 [3] and correct the ID switch with IDE [53] model. VVeRI-901 [23] only releases a trial version VVeRI-901-trial. We validate and compare the video object reID approaches on this trial version.

Evaluation Metric. We adopt the mean Average Precision (mAP) and the Cumulative Matching Characteristics (CMC) to evaluate the performance. The evaluation protocol is followed to BiCnet-TKS[19].

Training Details. For our implementation, we randomly choose 16 identities, and sample 4 sequences for each identity. For each sequence, we follow the restricted random sampling strategy [27], which divides each sequence into 8 chunks and randomly chooses one frame from each chunk. All video frames are resized to 256×128 after random data-augmentation (*i.e.,* random horizontal flipping, padding, random cropping and random erasing [15]). As for the optimizer, the SGD optimizer is employed and the learning rate is initialized

Table 1: The statistics of video object reID datasets.

Dataset	# ID	# Boxes	# Tracks	# Cams	#Frames
MARS	1,261	10,675,516	20,715	6	2∼920
MARS_DL	1,266	1,019,880	16,360	6	2∼920
PRID2011	178	38,466	354	2	5∼675
LSVID	3,772	2,982,685	14,943	15	60∼2533
iLIDS-VID	300	43,800	600	2	23∼192
VVeRI-901-trial	95	52,951	257	11	51∼462

as 0.01 with cosine learning rate decay. The total training epoch is set to 30. We set 3 shapes for multi-shape patch embedding to obtain diversity semantic representation: (1) 16×16, (2) 16×32, (3) 32×16.

4.2 Results on Video Person reID

In Tab. 2, we compare CAViT with state-of-the-arts on MARS and LSVID. CAViT achieves the best performance on all evaluation criteria. Tab. 2 shows the comparison on the two largest datasets (MARS and LS-VID) and Tab. 3 shows the comparison on the two small datasets (PRID-2011 and iLIDS-VID). In order to make a comparison with temporal shift based methods, we reproduce TSM with a ResNet50 backbone in video reID datasets. The Token shift module is reproduced by ourselves in video reID datasets. For a fair comparison, the token shift method is reproduced with the same pre-trained model (ViT_Base with 16x16 patch shape) and the same hyperparameters as CAViT.

CAViT *vs.* ViT baseline. We implement a strong video reID baseline model, which adapts ViT_Base [9] as the backbone, extracting features of all frames and compute the average feature for pedestrian retrieval. (1) CAViT improves ViT baseline over all six datasets. Particularly on LS-VID, CAViT obviously outperforms ViT baseline by 2.8%/3.9% mAP/rank-1. This is because that the misaligned problem in LS-VID is more seriousness than other datasets. (2) We also note that CAViT only achieves a 0.4% rank1 improvement on MARS. This is because that MARS has a lot of ID switch noise, which is caused by the tracking and detection algorithms. As shown in Tab. 4, after re-detection, CAViT achieves a 1.0% improvement on MARS_DL, even though ViT baseline has achieved high performance (94.6% rank1).

CAViT *vs.* 3D solutions. Existing joint learning solutions use 3D CNNs to jointly model the spatio-temporal clues. Compared with pure 3D CNN based methods, CAViT outperforms P3D [40] with 4.2%/2.2% mAP/rank-1 on MARS. Compared with temporal feature alignment method BiCnet-TKS, CAViT outperforms it by 1.2%/0.6% mAP/rank-1 on MARS, 4.1%/4.6% mAP/rank-1 on LSVID. Compared with temporal feature reconstructing method AP3D [13], CAViT outperforms it by 2.5%/1.2% mAP/rank-1 on MARS and 4.6% rank1 on iLIDS-VID. We argue that this is because 3D CNN is limited by the local receptor field of the convolutional network. Neither temporal alignment methods nor temporal reconstruction methods can solve the case of misalignment

Table 2: Comparison with state-of-the-arts on MARS [53], LS-VID [25] datasets. The methods are separated into two groups, the 2D neural network solutions (2D), and 3D neural network based solutions (3D).

Methods		Proc.	MARS		LS-VID	
			mAP	R-1	mAP	R-1
2D	MG-RAFA [50]	CVPR 20	85.9	88.8	-	-
	PhD [51]	CVPR 20	86.2	88.9	-	-
	AGRL [45]	TIP 20	81.9	89.5	-	-
	STGCN [47]	CVPR 20	83.7	90.0	-	-
	MGH [46]	CVPR 20	85.8	90.0	-	-
	RGTR [29]	AAA 21	84.0	89.4	-	-
	CTL [34]	CVPR 21	86.7	91.4	-	-
	GRL [35]	CVPR 21	84.8	91.0	-	-
	STRF [1]	ICCV 21	86.1	90.3	-	-
	PSTA [43]	ICCV 21	85.8	**91.5**	-	-
	DI [17]	ICCV 21	87.0	90.8	-	-
	STMN [10]	ICCV 21	84.5	90.5	69.2	82.1
	RFCnet [22]	PAMI 21	86.3	90.7	-	-
3D	I3D [5]	CVPR 17	83.0	88.6	33.9	51.0
	P3D [40]	ICCV 17	83.2	88.9	35.0	53.4
	IAUNet [21]	TNNLS 20	85.0	90.2	-	-
	M3D [26]	TPMAI 20	79.5	88.6	-	-
	TCLNet [20]	ECCV 20	85.1	89.8	-	-
	AP3D [13]	ECCV 20	85.1	90.1	-	-
	AFA [7]	ECCV 20	82.9	90.2	-	-
	STRF [1]	ICCV 21	86.1	90.3	-	-
	BiCnet-TKS [19]	CVPR 21	86.0	90.2	75.1	84.6
2D	TSM(R50) [32]	ICCV 19	81.8	88.6	66.0	78.3
	Token shift [7]	MM 21	86.6	90.2	68.7	80.4
	ViT baseline[9]	ARXIV 20	86.4	89.7	76.4	85.3
2D	**CAViT**	**Our work**	**87.2**	90.8	**79.2**	**89.2**

between frames well. Different with them, CAViT implements alignment of the entire frames, thus solving the misalignment well.

CAViT *vs.* 2D solutions. 2D solutions often apply 2D CNNs to model spatial clues and then use a temporal aggregation module(*i.e.,* LSTM, RNN, GCN, transformer) to merge the spatial representations. As we can see, DI is lower than CAViT by 0.2% on mAP in MARS and 1.3% rank1 in iLIDS-VID, while PSTA is lower than CAViT by 1.8% mAP on MARS and 1.8% rank1 on iLDS-VID. This is because lacking consideration of spatio-temporal interactions, they cannot take full advantage of the complementarity of adjacent frames.

CAViT *vs.* Temporal Shift Methods. TSM [32], token shift [49] and our CAViT use the temporal shift strategy for jointly modeling spatio-temporal clues with a 2D model. In Tab. 2, the performance gap is significant among these two methods and our CAViT. Specifically, on iLIDS-VID, CAViT outperforms

Table 3: Comparison with state-of-the-arts on PRID2011 [18], and iLIDS-VID [42] datasets. The methods are separated into two groups, the 2D neural network solutions (2D), and 3D neural network based solutions (3D).

Methods		Proc.	PRID-2011		iLIDS-VID	
			R-1	R-5	R-1	R-5
2D	MG-RAFA [50]	CVPR 20	95.9	99.7	88.6	98.0
	PhD [51]	CVPR 20	96.6	97.8	-	-
	AGRL [45]	TIP 20	94.6	99.1	84.5	96.7
	ADFD [52]	CVPR 19	93.9	99.5	86.3	97.4
	GLTR [25]	ICCV 19	95.5	100.0	86.0	98.0
	MGH [46]	CVPR 20	94.8	99.3	85.6	97.1
	RGTR [29]	AAAI 21	93.7	99.0	86.0	98.0
	GRL [35]	CVPR 21	96.2	99.7	90.4	98.3
	PSTA [43]	ICCV 21	95.6	98.9	91.5	98.1
	DI [17]	ICCV 21	-	-	92.0	98.0
3D	STRF [1]	ICCV 21	-	-	89.3	-
	M3D [26]	TPMAI 20	**96.6**	**100.0**	86.7	98.0
	TCLNet [20]	ECCV 20	-	-	86.6	-
	AP3D [13]	ECCV 20	-	-	88.7	-
	AFA [7]	ECCV 20	-	-	88.5	96.8
2D	TSM [32]	ICCV 19	87.6	93.5	69.3	81.3
	Token shift [7]	MM 21	91.1	95.5	86.0	98.0
	ViT baseline[9]	ARXIV 20	92.4	96.8	90.2	93.7
2D	**CAViT**	**Our work**	95.5	98.9	**93.3**	**98.0**

TSM by 30.0% rank1 and outperforms the token shift method by 7.3% rank1. We argue that TSM directly shifts the feature channel, which may aggravate the spatial misalignment among pedestrians. The performance of Token shift is close to CAViT on almost all datasets, except LSVID, where CAViT outperforms Token shift by 10.5% mAP and 8.8% rank1. This is because the misalignment is much more serious and the CLS token worsens this misalignment, making the performance of Token shift even worse than origin ViT model.

CAViT *vs.* Transformer-based Methods. Both of DI (Dense Interaction) [17] and our CAViT belong to the transformer based video reID methods. The difference is that DI applies the ResNet50 to extract spatial features and uses the transformer for temporal modeling, which is essentially a divide-and-conquer method instead of our joint modeling strategy. The proposed joint modeling method CAViT outperforms DI by 0.2% mAP on MARS and 1.3% rank1 on iLIDS-VID.

4.3 Results on Video Vehicle reID

We validate the proposed CAViT on the VVeRI-901-trial dataset. For a fair comparison, we also reproduce some widely used video representation learning methods(*e.g.*, AP3D, 3D Non-local and strong baseline of object reID) on the

Table 4: Comparison with state-of-the-arts on the MARS_DL dataset.

Methods		Proc.	MARS_DL	
			mAP	R-1
3D	TCLNet [20]	ECCV 20	85.4	91.0
	AP3D [13]	ECCV 20	86.5	91.3
	P3D-C [40]	ICCV 17	85.0	91.0
	C2D [24]	CVPR 19	86.2	91.4
2D	Non-Local [33]	ARXIV 19	85.8	90.8
	FT-WFT [39]	AAAI20	83.8	91.0
	DL+CF-AAN [37]	ARXIV 21	86.5	91.3
	TSM+ResNet [32]	ICCV 19	86.0	93.6
	Token Shift [49]	ACMM 21	90.1	94.9
	ViT baseline [9]	ARXIV 20	89.4	94.6
2D	**CAViT**	**Our work**	**90.5**	**95.6**

Table 5: Comparison with state-of-the-arts on the VVeRI-901-trial dataset.

Method		Proc.	VVeRI-901		
			mAP	R-1	R-5
3D	C2D [24]	CVPR 19	57.3	50.2	72.5
	NL3D [33]	ARXIV 19	60.5	55.0	77.5
	AP3D [13]	ECCV 20	61.2	52.5	75.0
	AP3D+NL3D [13]	ECCV 20	60.2	50.0	80.0
	BiCnet-TKS [19]	CVPR 21	50.8	41.3	70.4
2D	TSM [32]	ICCV 19	55.1	45.0	72.5
	BOT[38]	CVPRW 19	61.6	55.3	77.5
	SBS[15]	ARXIV 21	62.4	57.5	75.1
	ViT baseline[9]	ARXIV 20	62.7	52.5	84.0
	Token shift [49]	ICCV 19	**67.4**	57.5	80.0
2D	**CAViT**	**Our work**	65.6	**60.0**	**84.8**

VVeRI-901-trial dataset in Tab. 5. Compared with 3D solutions, CAViT outperforms AP3D by 4.4% mAP. Compared with similar temporal shift based methods, CAViT outperforms token shift by 4.8% rank-5 and outperforms TSM 15.0% on rank1.

4.4 Ablations Studies

Ablation of the Backbones & Multi-shape patch Embedding (MPE). In Tab. 6, to compare the performance of several popular backbones, we use different backbones of extracting spatial representations and apply an average pooling module for temporal aggregation. We can observe that, 2D backbones perform better than 3D backbones (*i.e.,* Timeformer, Swin_base 3D), since 3D backbones are troubled with misalignment. In addition, according to rank1 in Tab. 6, we can observe that: ViT_Base + MPE > ResNeSt101 = ViT_Base > ResNet101 > ResNeSt200 > Swin_base > ResNet50. Although the 32×16 patch

Table 6: The ablation study of Backbones & Multi-shape Patch Embedding.

Backbone	Patch shape			MARS	
	16×16	16×32	32×16	mAP	R-1
ResNet50 [14]				83.7	87.6
ResNet101 [14]				84.0	89.9
ResNeSt101 [48]				84.3	90.1
ResNeSt200 [48]				83.2	89.1
Swin_base [36]				83.6	88.4
Swin_base 3D [37]				68.3	81.4
ViT_Base 3D [2]				81.9	87.5
ViT_Base [9]	✓			86.4	89.7
ViT_Base [9]		✓		82.9	88.6
ViT_Base [9]			✓	83.0	87.8
ViT_Base [9]	✓	✓	✓	**86.8**	**90.6**

and the 16×32 patch are worse than 16×16 patch, the MPE which ensembles these three shapes achieves the best performance. This is because patches of different shapes focus on information of different granularity and directions.

Table 7: The ablation study of the different module in CAViT.

	MARS_DL		PRID-2011	
	mAP	R-1	R-1	R-5
ViT baseline	89.4	94.6	92.4	96.8
+ MPE	90.0	94.8	93.8	97.7
+ TSA	90.2	95.3	94.6	98.0
+ RPE	**90.5**	**95.6**	**95.5**	**98.9**

Ablation of TSA & RPE. To denote the effectiveness of Temporal Shift Attention (TSA) moudle and Residual Position Embedding (RPE) moudle, we implement ablation experiments on MARS-DL and PRID-2011 in Tab. 7. Compared with the ViT model with multi-shape patch embedding, TSA achieves 0.2%/0.5% mAP/rank1 increment on MARS_DL. With RPE, TSA improves 0.5%/0.8% mAP/rank1 on MARS_DL and improves 1.7%/1.2% rank1/rank5 on PRID-2011. This indicates that TSA notices temporal saliency clues, under the guidance of RPE.

Attention map visualization. The normalized attention maps of MPE are visualized in Fig. 4. (1) For spatial clues, according to this figure, different shape of MPE has different attention regions. MPE helps CAViT pay attention to a variety of granularity and directions and obtain more diverse spatial representations. (2) For temporal clues, the deeper the network layer, the more attention pays on adjacent frames. It also indicates that CAViT learns spatial clues in shallow layers and implements temporal alignment in deep layers.

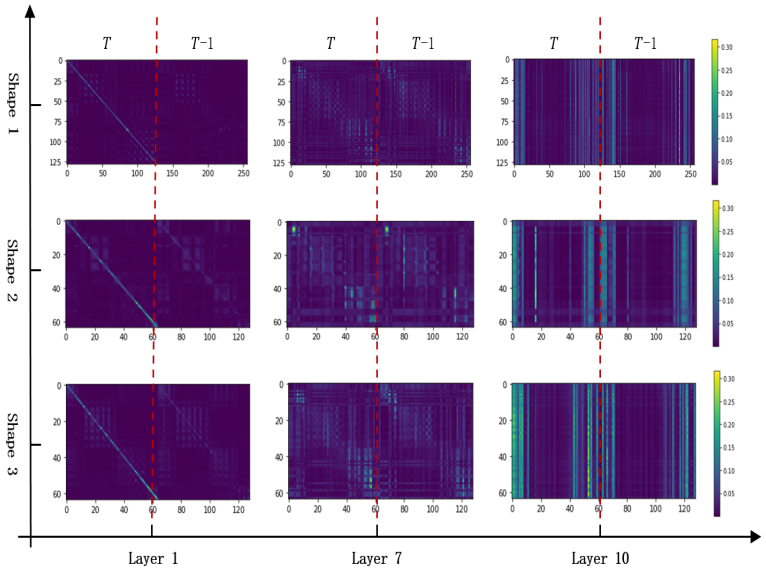

Fig. 4: Attention map visualization for MPE. (1)the first row belongs to the 16x16 patch, while the second and the third rows belong to the 16x32 patch and the 32x16 patch, respectively. (2) For each sub-figure, the left half part is the attention weight of the intra-frame, while the right part is the attention weight of the adjacent frame. (3) Different columns represent the attention map at different layers.

5 Conclusion

This paper proposes a contextual alignment vision transformer (CAViT) for the video object re-identification, which contains a multi-shape patch embedding module (MPE) and a Temporal Shift Attention (TSA) module. The former obtains diversity semantic embedding for spatial alignment in the pedestrian matching process, while the latter applies a 2D solution for jointly modeling spatio-temporal clues. We also introduce a residual position embedding (RPE) to guide the temporal shift attention in focusing on temporal saliency clues. Experimental results on five video pedestrian reID datasets and one video vehicle reID dataset demonstrate the superiority of the proposed CAViT over state-of-the-art methods.

Acknowledgements

This research was supported by the National Key R&D Program of China under Grant No.2020YFC2003901, Chinese National Natural Science Foundation Projects 61876178, 61872367, 61976229, 62176256, 62106264 and the InnoHK program.

References

1. Aich, A., Zheng, M., Karanam, S., Chen, T., Roy-Chowdhury, A.K., Wu, Z.: Spatio-temporal representation factorization for video-based person re-identification. In: ICCV (2021) 4, 10, 11
2. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? arXiv preprint arXiv:2102.05095 (2021) 13
3. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020) 8
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020) 4
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017) 10
6. Chen, C.F., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. arXiv preprint arXiv:2103.14899 (2021) 4
7. Chen, G., Rao, Y., Lu, J., Zhou, J.: Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In: ECCV (2020) 10, 11
8. Dehghan, A., Modiri Assari, S., Shah, M.: Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In: CVPR (2015) 8
9. 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. arXiv preprint arXiv:2010.11929 (2020) 2, 4, 7, 9, 10, 11, 12, 13
10. Eom, C., Lee, G., Lee, J., Ham, B.: Video-based person re-identification with spatial and temporal memory networks. In: ICCV (2021) 10
11. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV (2019) 3
12. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE TPAMI (2009) 8
13. Gu, X., Chang, H., Ma, B., Zhang, H., Chen, X.: Appearance-preserving 3d convolution for video-based person re-identification. In: ECCV (2020) 3, 9, 10, 11, 12
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 13
15. He, L., Liao, X., Liu, W., Liu, X., Cheng, P., Mei, T.: Fastreid: a pytorch toolbox for real-world person re-identification. arXiv preprint arXiv:2006.02631 (2020) 8, 12
16. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: Transformer-based object re-identification. arXiv preprint arXiv:2102.04378 (2021) 4
17. He, T., Jin, X., Shen, X., Huang, J., Chen, Z., Hua, X.S.: Dense interaction learning for video-based person re-identification supplementary materials. Identities 4, 10, 11
18. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person Re-Identification by Descriptive and Discriminative Classification. In: SCIA 8, 11
19. Hou, R., Chang, H., Ma, B., Huang, R., Shan, S.: Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In: CVPR (2021) 3, 8, 10, 12
20. Hou, R., Chang, H., Ma, B., Shan, S., Chen, X.: Temporal complementary learning for video person re-identification. In: ECCV (2020) 10, 11, 12

21. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Iaunet: Global context-aware feature learning for person reidentification. IEEE TNNLS (2020) 10
22. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Feature completion for occluded person re-identification. IEEE TPAMI (2021) 3, 10
23. Jianan Zhao, Fengliang Qi, G.R., Xu, L.: Vveri-901: Video vehicle re-identification dataset (2020), https://www.graviti.cn/open-datasets/VVeRI901 8
24. Li, C., Zhong, Q., Xie, D., Pu, S.: Collaborative spatiotemporal feature learning for video action recognition. In: CVPR (2019) 12
25. Li, J., Wang, J., Tian, Q., Gao, W., Zhang, S.: Global-local temporal representations for video person re-identification. In: ICCV (2019) 8, 10, 11
26. Li, J., Zhang, S., Huang, T.: Multi-scale 3d convolution network for video based person re-identification. In: AAAI (2019) 3, 10, 11
27. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: CVPR (2018) 8
28. Li, S.Z.: Markov random field modeling in image analysis. Springer Science & Business Media (2009) 5
29. Li, X., Zhou, W., Zhou, Y., Li, H.: Relation-guided spatial attention and temporal refinement for video-based person re-identification. In: AAAI (2020) 10, 11
30. Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., Wu, F.: Diverse part discovery: Occluded person re-identification with part-aware transformer. In: CVPR (2021) 4
31. Liao, S., Shao, L.: Transformer-based deep image matching for generalizable person re-identification. NeurIPS Workshops (2021) 4
32. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: ICCV (2019) 10, 11, 12
33. Liu, C.T., Wu, C.W., Wang, Y.C.F., Chien, S.Y.: Spatially and temporally efficient non-local attention network for video-based person re-identification. arXiv preprint arXiv:1908.01683 (2019) 12
34. Liu, J., Zha, Z.J., Wu, W., Zheng, K., Sun, Q.: Spatial-temporal correlation and topology learning for person re-identification in videos. In: CVPR (2021) 10
35. Liu, X., Zhang, P., Yu, C., Lu, H., Yang, X.: Watching you: Global-guided reciprocal learning for video-based person re-identification. In: CVPR (2021) 10, 11
36. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. ICCV (2021) 4, 13
37. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. arXiv preprint arXiv:2106.13230 (2021) 8, 12, 13
38. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: CVPR Workshops (2019) 12
39. Pathak, P., Eshratifar, A.E., Gormish, M.: Video person re-id: Fantastic techniques and where to find them. arXiv preprint arXiv:1912.05295 (2019) 12
40. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3d residual networks. In: ICCV (2017) 3, 9, 10, 12
41. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV (2015) 3
42. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: ECCV (2014) 8, 11
43. Wang, Y., Zhang, P., Gao, S., Geng, X., Lu, H., Wang, D.: Pyramid spatial-temporal aggregation for video-based person re-identification. In: ICCV (2021) 4, 10, 11

44. Weng, X., Kitani, K.: Learning spatio-temporal features with two-stream deep 3d cnns for lipreading. arXiv preprint arXiv:1905.02540 (2019) 3
45. Wu, Y., Bourahla, O.E.F., Li, X., Wu, F., Tian, Q., Zhou, X.: Adaptive graph representation learning for video person re-identification. IEEE TIP (2020) 10, 11
46. Yan, Y., Qin, J., Chen, J., Liu, L., Zhu, F., Tai, Y., Shao, L.: Learning multi-granular hypergraphs for video-based person re-identification. In: CVPR (2020) 10, 11
47. Yang, J., Zheng, W.S., Yang, Q., Chen, Y.C., Tian, Q.: Spatial-temporal graph convolutional network for video-based person re-identification. In: CVPR (2020) 4, 10
48. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., et al.: Resnest: Split-attention networks. arXiv preprint arXiv:2004.08955 (2020) 13
49. Zhang, H., Hao, Y., Ngo, C.W.: Token shift transformer for video classification. In: ACM MM (2021) 10, 12
50. Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In: CVPR (2020) 10, 11
51. Zhao, J., Qi, F., Ren, G., Xu, L.: Phd learning: Learning with pompeiu-hausdorff distances for video-based vehicle re-identification. In: CVPR (2021) 10, 11
52. Zhao, Y., Shen, X., Jin, Z., Lu, H., Hua, X.s.: Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In: CVPR (2019) 11
53. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: Mars: A video benchmark for large-scale person re-identification. In: ECCV (2016) 8, 10
54. Zhou, Z., Huang, Y., Wang, W., Wang, L., Tan, T.: See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In: CVPR (2017) 3
55. Zhu, K., Guo, H., Zhang, S., Wang, Y., Huang, G., Qiao, H., Liu, J., Wang, J., Tang, M.: Aaformer: Auto-aligned transformer for person re-identification. arXiv preprint arXiv:2104.00921 (2021) 4