VID-Trans-ReID: Enhanced Video Transformers for Person Re-identification

Aishah Alsehaim[1,2]

Toby P. Breckon[1]

[1] Department of Computer Science
Durham University, UK

[2] Department of Computer Science
Princess Nourah Bint Abdulrahman
University, SA

Abstract

Video-based person Re-identification (Re-ID) has received increasing attention recently due to its important role within surveillance video analysis. Video-based Re-ID expands upon earlier image-based methods by extracting person features temporally across multiple video image frames. The key challenge within person Re-ID is extracting a robust feature representation that is invariant to the challenges of pose and illumination variation across multiple camera viewpoints. Whilst most contemporary methods use a CNN based methodology, recent advances in vision transformer (ViT) architectures boost fine-grained feature discrimination via the use of both multi-head attention without any loss of feature robustness. To specifically enable ViT architectures to effectively address the challenges of video person Re-ID, we propose two novel modules constructs, Temporal Clip Shift and Shuffled (TCSS) and Video Patch Part Feature (VPPF), that boost the robustness of the resultant Re-ID feature representation. Furthermore, we combine our proposed approach with current best practices spanning both image and video based Re-ID including camera view embedding. Our proposed approach outperforms existing state-of-the-art work on the MARS, PRID2011, and iLIDS-VID Re-ID benchmark datasets achieving 96.36%, 96.63%, 94.67% rank-1 accuracy respectively and achieving 90.25% mAP on MARS.

1 Introduction

Video based person re-identification (Re-ID) is a popular research area within computer vision gaining increasing attention due to a wide range of potential applications, such as intelligent video surveillance and automated security systems. Video person Re-ID refers to the task of matching a person in a query surveillance video, to the same person within other videos from multiple non-overlapping camera views. However, this poses a very challenging problem due to variations in human pose, occlusion, differing camera viewpoints, illumination and background scene clutter. In general, unlike its single-frame image-based counterpart [9, 38, 48, 70], video-based Re-ID benefits from rich multi-frame, temporal information to address the task of cross-video instance matching. Video-based Re-ID has also benefited from the significant development of deep learning methods by building differing structural approaches that learn discriminative and robust deep features of person subjects in a video [9, 10, 27, 28, 37, 55, 65].

Of late, CNN-based methods have dominated and achieved remarkable success in both image-based [23, 35, 43, 68] and video-based [9, 10, 27, 28, 37, 55, 65] person Re-ID with

CNN-based feature extraction achieving superior results across most deep feature based methods [2, 10, 27, 28, 37, 55, 65]. However, more recently vision transformer (ViT) architectures [25, 34, 45, 58] have shown very notable progress across a number of image understanding tasks, including object detection [6, 16, 34, 72], image segmentation [34], and image classification [25, 34]. Transformers with multi-head attention and without any down sampling operations potentially offer more effective frame-level feature extraction for video person Re-ID tasks where it is imperative to extract fine-grained on-person feature details. However within CNN based methods, such fine-grain features - that can potentially boost the re-identification process - are often lost in the feature extraction process due to the prevalence of multiple pooling and convolutional operations over varying strides. In contrast, transformer based architectures tend to retain all of the visual information that is required to boost person re-identification whilst additionally capturing detailed long range feature dependencies. In this way, the multi-head attention within a transformer architecture has the potential to capture long range feature dependencies, in contrast to CNN based models that extract small discriminative regions, necessitating additional attention blocks [4, 30, 54, 64]. To date, very recent prior work using such ViT architectures to address the Re-ID problem are limited to the non-temporal problem of image-based Re-ID [16, 71] or alternatively do not jointly consider both the spatial and temporal relation of the video Re-ID task within a single transformer [61].

In this paper, we introduce an enhanced vision transformer (ViT) as a novel backbone architecture for frame-level feature extraction within video based Re-ID. Moreover, we incorporate additional novel modules to address the specific person Re-ID challenges of occlusion, pose variation and camera view variation. Inspired by positional embedding in existing vision transformers, we additionally add a learnable camera ID (camera view) embedding to our patch embeddings to address camera view variation in the Re-ID task.

After extracting frame-level features using our ViT backbone, we introduce our novel **Temporal Clip Shift and Shuffle (TCSS)** module to shift features between frames and jointly shuffle the frame feature order. Subsequently, the resulted video features are more robust to occlusion and pose variation. In our method, we generate video-level features both globally and locally. To extract robust global video-level features we aggregate frame-level features extracted by the ViT using temporal attention. Additionally, we propose an efficient novel **Video Patch Part Feature (VPPF)** module to extract local video features across multiple video frames. VPPF ensures that frame patches with the same in-frame position are consistently used to generate local features that would be otherwise undiscoverable via global feature extraction alone. Our subsequent results show that this dual use of both local and global video-level features significantly boosts the re-identification performance (Section 4). In summary, the main contributions of this paper are as follows:

- An enhanced vision transformer (ViT) based architecture for video person Re-ID, incorporating both global video-level features and local video patch part features.
- Our novel Temporal Clip Shift and Shuffle (TCSS) and Video Patch Part Feature (VPPF) modules that are subsequently experimentally shown to provide robust fine-grained feature extraction to boost overall video person Re-ID performance.
- Our experimental results, based on the use of this enhanced vision transformer (ViT) architecture, that achieve state of the art performance on the MARS (96.36%), PRID2011 (96.63%), and iLIDS-VID (94.67%) for rank-1 accuracy.

2 Related Work

In video person Re-ID the primary aim is to extract robust person feature representations avoiding all auxiliary spatial or temporal scene distractors. All existing methods mainly focus on efficiently extracting rich spatial-temporal features, most recently leveraging CNN networks [2, 3, 17, 32, 51] to extract per-frame appearance representation followed by varying methods to aggregate temporal information. Some approaches then perform spatial-temporal average or max pooling [7, 14, 33] on video frames to generate video-level features. Attention models [29, 43, 69] are rapidly becoming the most common strategy to aggregate initial video frame features. Attention based aggregation focuses on selecting the most informative frames, while average and max pooling treats all the frames equally. Whilst many techniques use optic flow [7, 37, 55], many person Re-ID applications require real-time performance precluding the use of such computationally expensive techniques. Alternatively, recurrent CNN are also explored to capture the temporal structure and aggregate temporal features within videos [7, 55, 69].

In recent work, 3D convolution has been adopted for video feature learning in video person Re-ID, as it directly extracts spatial-temporal features [41, 46]. Multi-scale 3D (M3D) CNN [27] uses 3D convolutions to extract spatial-temporal features but requires a significantly larger number of parameters to be optimised resulting in both additional computational complexity and an increased memory footprint for both training and inference.

More recently, the use of graph neural networks for video Re-ID has been introduced in [57], where two separate graph networks for spatial and temporal features are created and jointly optimised to extract video spatial-temporal features. Yan et al. [56] use multiple hypergraphs representing different granularities, wherein graph nodes are constructed according to part-based body features. Spatial-temporal features are then aggregated via a graph convolution network.

Conversely, vision transformers (ViT) are gaining significantly more traction of late [25, 34, 45, 58] and, with their multi-head attention and strong fine-grain feature retention, yield a highly suitable feature extractor for video person Re-ID. However, video person Re-ID differs from other computer vision tasks by jointly facing the combined challenges of human pose variation, occlusion, differing camera viewpoints, illumination, and background scene clutter simultaneously. To these ends, the very recent work of Zang et. al [61] proposed a multi-direction and multi-scale Pyramid in Transformer (PiT) that looks at each frame without division, with vertical patch division and horizontal patch division in addition to the patch-based division strategies to explore the fine-grained features. By contrast, we employ a classical vision transformer architecture (ViT) to provide frame-level feature with only patch-based division strategies. This is enhanced with our Temporal Clip Shift and Shuffle (TCSS) and Video Patch Part Feature (VPPF) modules to subsequently address these combined challenges of video person Re-ID via the learning of more robust discriminative video features within a ViT based formulation.

3 Methodology

We describe our video Re-ID methodology by first outlining our approach of frame-level feature extraction and subsequently introducing our novel Temporal Clip Shift and Shuffle (TCSS) and Video Patch Part Feature (VPPF) modules. Finally, we detail our loss function optimization strategy for the overall ViT based network architecture proposed.

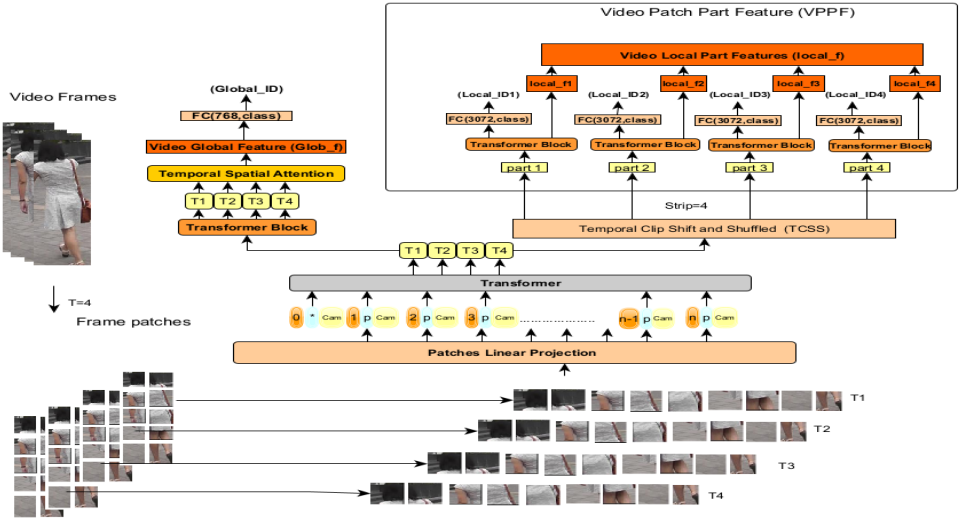


Figure 1: Our VID-Trans-ReID architecture with VPPF and TCSS modules shown.

3.1 Feature Extraction

Consider a surveillance video V_i with number of frames as $V_i = \{I_1, I_2, .., I_k\}$. We split frame $I_j \in \mathbb{R}^{H\times W\times C}$ where H, W, C are the height, width and number of channels respectively, into n equal size patches as $I_j \in V_i$ and $I_j = \{I^j_{p_1}, I^j_{p_2}, .., I^j_{p_n}\}$ such that our patches overlap to avoid information loss between patches [16, 49, 60]. The overlap region for patch size $P \times P$ and step size s is $(P-s)P$. The number of patches for a frame of size $H \times W$ can be thus calculated as:

$$n = \lfloor \frac{H+s-P}{s} \rfloor \times \lfloor \frac{W+s-P}{s} \rfloor \tag{1}$$

Subsequently, in video person Re-ID each frame I_j with n patches is prepended with a learnable class embedding I_{cls} as follows:

$$I_j = [I^j_{cls}; \mathcal{F}(I^j_{p1}); \mathcal{F}(I^j_{p2}); \cdots ; \mathcal{F}(I^j_{p(n)})] \tag{2}$$

where I^j_{cls} is frame class token to represent the global feature of the frame and $\mathcal{F}$ is the mapping function that project patches to D dimensional space. The spatial information is preserved within the transformer architectural formulation by adding a learnable positional embedding pos, yielding:

$$I_j = I_j + pos\ , \quad \text{where } pos \in \mathbb{R}^{(n+1)\times D}. \tag{3}$$

Given the multi-camera nature of video person Re-ID, we add additional learnable camera embedding to represent the camera ID of a given view, inspired by positional embedding within the baseline transformer architecture. Prior work indicates the effectiveness of this lightweight learnable embedding for learning invariant non-visual features [16, 40]. Camera embedding is a learnable 1-D embedding cam, where $cam \in \mathbb{R}^{C\times D}$ and C denotes the number of camera views within the dataset. In contrast to position embedding, all patches in a given video will carry the same cam value: if video V_i is recorded by camera m, then cam_m is the camera embedding for all patches in this video. The video frame input passed to the transformer layers can thus be expressed as:

$$I_j = I_j + \lambda\, pos + (1-\lambda) cam_m \tag{4}$$

where I_j is the frame patch prepended by frame class token, *pos* denotes position embedding for each patch, cam_m is the camera embedding for a frame recorded by camera m and λ balances the weight of the positional embedding and camera embedding.

3.2 Video Level Features

To generate effective video-level features we jointly use both a global branch to extract global video features (Sec. 3.2.1) and a local branch to extract fine-grained features using our novel Temporal Clip Shift and Shuffle (TCSS) and Video Patch Part Feature (VPPF) modules (Sec. 3.2.2 / 3.2.3). An overview of our proposed architecture is shown in Figure 1.

3.2.1 Global Video Features

In the global branch the model learns to produce clip-level features, C, at the training stage by choosing random frames, T, from the tracklet. At inference time, all of the frames in a tracklet are used to produce the video-level feature by dividing the video V_i into several clips as V_i=$[C_1, C_2, ..., C_m]$, where each clip C_i has T frames C_i=$[I^1, I^2, ..I^T]$, and T is the number of selected frames to train the network. A transformer network is used to extract features at the frame-level. These features are then aggregated in the global branch to clip-level features using spatio-temporal attention [14]. Here our spatio-temporal attention is a 2D convolution with an input dimensionality of 768 mapped to a 256 dimensional output, followed by a 1D temporal convolution on the frame-level features to generate temporal attention s_i^t. The final frame attention score a_i^t is calculated using $softmax(\cdot)$ [69]. At the end of this branch, one fully connected layer applied to clip C_i features to predicts the person ID (*Global_ID* in Figure 1).

3.2.2 Temporal Clip Shift and Shuffle (TCSS)

To deliver more discriminative fine-grained video features we additionally propose a local feature branch in parallel with earlier global branch (see Figure 1). This local branch extracts fine-grained features using our novel Temporal Clip Shift and Shuffle (TCSS) and Video Patch Part Feature (VPPF) modules. Here, frame features T are extracted by the transformer and concatenated at the patch level to form clip features, C_i as follows:

$$\mathcal{C}_i = [P_0\{I^1_{p_0}, I^2_{p_0}, \cdots, I^T_{p_0}\}; \cdots; P_n\{I^1_{p_n}, I^2_{p_n}, \cdots, I^T_{p_n}\}] \tag{5}$$

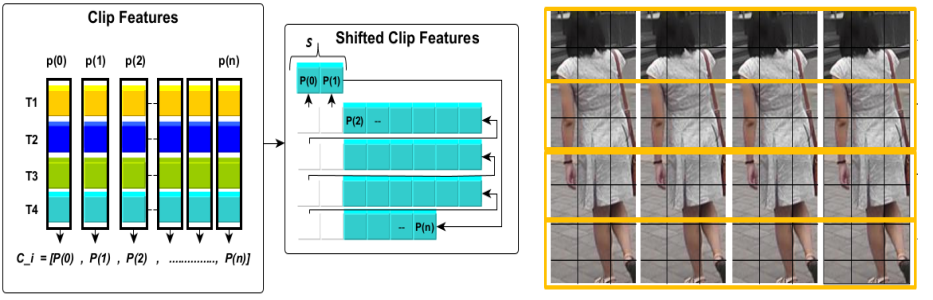
As shown in Figure 2a (Clip Features), each clip C_i is passed to the Temporal Clip Shift and Shuffle (TCSS) module that takes T frame-level with n patch features from clip C_i and then concatenates these features at the patch level by connecting all the patches at the same position across different adjacent frames. These clip features are then shifted on a patch-wise basis by S steps (Figure 2a, Shifted Clip Features) as follows:

$$\mathcal{C}_i^{sh} = [\{C_i^{P_{(0+S)}}\}^0_{T*D}; \cdots; \{C_i^{P_{(n)}}\}^0_{T*D}; \{C_i^{P_{(0)}}\}^0_{T*D}; \cdots; \{C_i^{P_{(s)}}\}^0_{T*D}] \tag{6}$$

where $\mathcal{C}_i^{sh}$ is the clip features shifted by S steps. Inspired by He et al. [16] and shuffleNet [62], where image shuffling boosts the fine-grained feature extraction, we apply a further shuffling layer to clip-level shifted features C_i. Our shuffling process is performed at the patch level by dividing it into two groups and then shuffling.

3.2.3 Video Patch Part Feature (VPPF)

Part-based localised part features have been successfully used in earlier CNN-based methods to extricate fine-grained features for person Re-ID [36, 43, 47, 59]. Inspired specifically by



(a) Temporal Clip Shift.

(b) Video Patch Part Feature (VPPF) with 4 strips.

Figure 2: Temporal Clip Shift and Shuffle (TCSS) and Video Patch Part Feature (VPPF).

local strip features in CNN-based methods [36, 43, 47, 59], we propose novel Video Patch Part Feature (VPPF). VPPF extracts fine-grained video features by dividing image patch features into horizontally aligned strips from which we extract strip features, not only from each frame but from all the frames T within the video clip. In our method we use strip length, $strip = 4$, as shown in Figure 2b. These strip features are then passed to a single transformer layer to generate video local part features (($local\_f1$, $local\_f2$, $local\_f3$, $local\_f4$) in Figure 1), which are subsequently passed to multiple fully connected layers equal to the number of strips to predict per-strip person ID. At inference time, we use a combination of both Video Global Features generated by global branch and Video Local Part Features generated by the Video Patch Part Feature (VPPF) module to represent the features of a person in a query or gallery video (see Figure 1).

3.3 Model Optimisation

Our two branches, video global features and video local part features are optimised using a combined loss function comprising label smoothing cross entropy loss $\mathcal{L}_{ID}$ [44] , triplet loss $\mathcal{L}_{triple}$ [18], $\mathcal{L}_{E_{att}}$ [39] and center loss [53] as follows:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{ID}(Glob_{ID}) + \mathcal{L}_{triple}(Glob_f) + \beta \times \mathcal{L}_{center}(Glob_f) + \mathcal{L}_{E_{att}} \\ &+ \frac{1}{stripes} \sum_{j=1}^{stripes} (\mathcal{L}_{ID}(Local_{ID}^j) + \mathcal{L}_{triple}(local_f^j) + \mathcal{L}_{center}(local_f^j))\end{aligned} \tag{7}$$

where are $Glob_f$ is video-level feature extracted by global branch, $local_f$ is video local part features extracted by VPPF, ($Glob_{ID}$ and $Local_{ID}$) are the predicted ID labels and $stripes$ is the number of stripes used within VPPF (Figure.1, Figure.2b) and beta is β = 0.00005. Within this formulation (Eqn. 7), we use commonplace person Re-ID losses and smooth both cross entropy loss [44] and triplet loss [18] to both support more robust discriminative feature and to pool samples by similarity within the resulting feature embedding space. We also include center loss [53] with the aim of further regularising the inter instance distances within the feature embedding space on a class-wise basis.

4 Evaluation

We evaluate our approach using three established benchmark video person Re-ID datasets (MARS, PRID2011, iLIDS-VID). MARS [66] is the largest video person Re-ID benchmark

that consists of 1261 identities and 20,715 tracklets under 6 camera views with bounding boxes from the DPM detector [13] and GMMCP tracker [11]. PRID2011 [19] contains 400 tracklets of 200 identities from two cameras varying from 5 to 675 frames. iLIDS-VID [50], contains 600 tracklets of 300 identities from 2 cameras varying from 23 to 192 frames.
For statistical evaluation, we use the Cumulative Match Characteristic (CMC) and mean Average Precision (mAP) metrics. CMC is used to evaluate model performance in identifying the correct identity within the top-k ranked matches (reporting Rank-1 accuracy for each dataset). The mAP metric is used to evaluate model performance across multi-shot re-identification datasets such as MARS [66].

4.1 Implementation Details

All video frames are resized to 256×128 and padded with 10 zero-valued pixels. Each training example is flipped horizontally with random erasing [67] using 0.5 probability. Normalised RGB pixel triplets are rescaled via division by $(0.5, 0.5, 0.5)$ and normalized via subtraction of $(0.5, 0.5, 0.5)$, respectively (following [25]). The batch size is set to 32 with 4 videos for each ID. Our optimizer is Stochastic Gradient Descent (SGD) with learning rate equal to 0.008 with cosine learning rate decay ($momentum = 0.9$, $weightdecay = 1e-4$). The initial weights of ViT [25] for the MARS [66] dataset are pre-trained on ImageNet-21K and pretrained on MARS [66] for iLIDS and PRID. Our model is trained for 120 epochs and trained to generate clip features, where each clip consists of $T = 4$ frames chosen randomly from the video for each ID. We use $\lambda = 0.25$ for the transformer input in order to give more weight to camera embedding, to better capture change in video view whilst positional embedding remains the same in all frames, however.

4.2 State-of-the-Art Comparison

Our experiments show that the use of our approach improves Re-ID performance against leading contemporary approaches by over a 5% margin for CMC Rank-1 on the most challenging Re-ID dataset MARS [66] (see Table 1). With the smaller Re-ID datasets (iLIDS-VID [50], PRID2011 [19]), pre-trained on MARS [66], our method improves CMC Rank-1 performance on iLIDS-VID [50] by 2%+. Furthermore, we narrowly improve CMC for PRID2011 [19] whilst we also demonstrate state of the art mAP accuracy for the MARS dataset (90.25%) (see Table 1).

These experiments support our hypothesis that a transformer based video Re-ID approach can indeed outperform contemporary CNN based methods since there is no loss of any information through convolutional-based pooling and stride. Furthermore, the long range feature dependencies using multi-head attention in transformers deliver accurate detailed fine-grained information that maximises person Re-ID performance.

5 Ablation Study

The effectiveness of our proposed Temporal Clip Shift and Shuffle (TCSS) and Video Patch Part Feature (VPPF) modules at extracting robust person features is illustrated in Tables 2 and 3. In Table 2 we can see that removal of the TCSS module degrades both CMC Rank-1 and mAP by approximately -3% on MARS [66]. Similarly, we show that without the support of VPPF, our model loses approximately -2% in Rank-1 accuracy and about -1% in mAP on MARS [66].

We also study the effect of removing *Cam*, the learnable camera embedding. Table 2 shows that a drop in performance of approximately 5% in Rank-1 and about 9% in mAP for

Methods	Publication	MARS [66] Rank-1 (mAP)	iLIDS-VID [50]	PRID2011 [19]
Att-Driven [65]	CVPR 2019	87.0 (78.2)	86.3	-
VRSTC [20]	CVPR 2019	88.5 (82.3)	83.4	-
Co-Segment [42]	ECCV 2019	84.9 (79.9)	77.8	-
GLTR [26]	ICCV 2019	87.02 (78.47)	86	95.50
M3D [27]	IEEE-T IP 2020	88.63 (79.46)	86.67	96.60
TACAN [28]	WACV 2020	89.1 (84.0)	88.9	95.3
AP3D[15]	ECCV 2020	90.1(85.1)	-	-
AFA [8]	ECCV 2020	90.2(82.9)	88.5	-
TCLNet [21]	ECCV 2020	89.8(85.1)	86.6	-
MGH [56]	CVPR 2020	90(85.8)	85.6	94.8
MG-RAFA [63]	CVPR 2020	88.8(85.9)	88.6	95.9
STGCN [67]	CVPR 2020	89.95 (83.70)	-	-
Co-Aware[64]	IEEE TIFS 2021	88.2 (84.1)	85.8	92.2
HMN [62]	IEEE TCS VT 2021	89.0 (88.80)	-	-
SANet[4]	IEEE TCS VT 2021	91.2 (86.0)	-	-
STMN [12]	ICCV 2021	90.5 (84.5)	-	-
STRF [1]	ICCV 2021	90.3 (86.10)	89.30	-
DenseIL [17]	ICCV 2021	90.8 (87.0)	92.0	-
PSTA [51]	ICCV 2021	91.5 (85.8)	91.5	95.6
SSN3D [24]	AAAI2021	90.1(86.2)	88.9	-
CTL [31]	CVPR2021	91.40(86.70)	89.70	-
Watching You [32]	CVPR 2021	91.0 (84.8)	90.4	96.2
BiCnet-TKS [22]	CVPR 2021	90.2 (86.0)	-	-
PiT [61]	IEEE TII 2022	90.22 (86.80)	92.07	-
SINet [6]	CVPR 2022	91.0 (86.2)	92.5	96.5
ViT [25] (baseline)		90.68(78.61)	38.67	74.16
VID-Trans-ReID (**ours**)	–	**96.36 (90.25)**	**94.67**	**96.63**

Table 1: Video person Re-ID: state-of-the-art comparison

MARS [66] with camera embedding removed. Similarly, the removal of camera embedding decreases Rank-1 by 8% and 3% in iLIDS[50] and PRID [19] respectively. Within person Re-ID, information performance between neighbor patches is conveyed via the use of patch overlapping within the Re-ID process is also of great importance (Table 2 / 3). Table 2 and 3 similarly show that the use of pure transformer settings (without patch overlapping) results in significant degradation on both Rank-1 and mAP.

Method	Rank-1	mAP
ours *without* TCSS	93.80	86.26
ours *without* VPPF	94.62	88.70
ours *without* camera embedding	91.60	80.73
ours *without* patch overlapping	91.44	78.55
VID-Trans-ReID (**our full method**)	**96.36**	**90.25**

Table 2: Ablation results: MARS.

Method (with Rank-1 reported only)	iLIDS	PRID
ours *without* TCSS	93.33	94.63
ours *without* VPPF	89.0	94.38
ours *without* camera embedding	86.67	93.26
ours *without* patch overlapping	89.33	95.51
VID-Trans-ReID (**our full method**)	**94.67**	**96.63**

Table 3: Ablation results: iLIDS & PRID.

In addition, we also consider the use of our novel TCCS and VPPF modules within other leading contemporary transformer architectures when applied to video person Re-ID, such as Swin transformer [34] and the Focal transformer [58] with their original implementation, without the use of camera embedding and overlapping patches. As most transformer architectures are suitable for uniformly square frames, whilst in person Re-ID most of the subject frames are non-uniform (following the aspect ratio of the human body), we consider performance based on both the use of non-uniform (Table 4) and uniform (Table 5) frame size. We observe that our method, using a ViT architecture, [25] outperforms Swin [34] and Focal

[53]) by 3% even with the addition of our novel TCSS and VPPF modules to each of these architectures. In both cases the addition of these modules to Swim and Focal only make a marginal performance improvement that remains significantly lesser when compared to our own approach (Table 4). Whilst the use of uniform frame size is shown to further improve performance (for Swin, Focal), the use of TCCSS and VPPF provide further improvement across both (Table 4) and our approach is still shown to outperform on the challenging MARS dataset (Table 4, Table 5 - mAP).

Methods	Rank-1	mAP
Swin (*baseline*) [34]	78.79	53.38
Swin + TCSS + VPPF	80.30	54.17
Focal (*baseline*) [53]	92.36	78.87
Focal + TCSS + VPPF	93.08	81.14
VID-Trans-ReID (**ours**)	**96.36**	**90.25**

Table 4: Non-uniform frame size: MARS.

Methods	Rank-1	mAP
Swin + TCSS + VPPF	**94.53**	81.86
Focal + TCSS + VPPF	93.08	81.14
VID-Trans-ReID (**ours**)	93.03	**84.96**

Table 5: Uniform frame size (224×224): MARS.

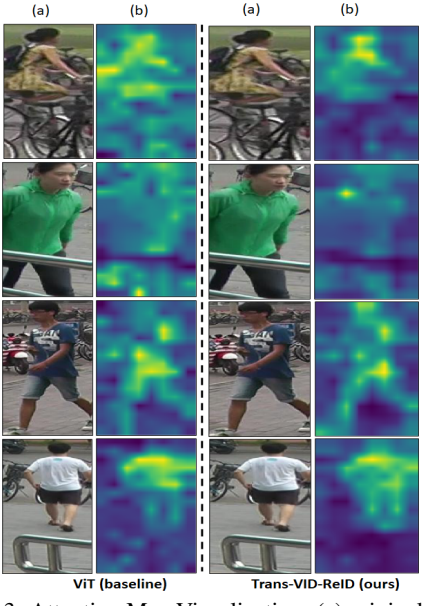


Figure 3: Attention Map Visualization: (a) original input samples, (b) attention maps - ViT vs. VID-Trans-ReID (ours).

Qualitatively we can also show that our proposed method extracts more focused person related features from the image and ignores any unrelated scene clutter more effectively. Fig. 3 clearly demonstrates the difference between using an unmodified ViT architecture [25] and our method with the additional TCSS and VPPF models added that result in a much more distinct separation of the on person features (via the attention map) from those of either the background or interacting objects such as the bicycle (Fig. 3).

Figure 4: Rank-10 results: VID-Trans-ReID (ours) & ViT (baseline) (***green*** = *correct*, ***red*** = *incorrect*)

If we consider the first row of Fig. 3, our method only captures on-person features via the use of attention while bicycle features are ignored, resulting in improved Re-ID against other videos of the same person without the bicycle present. This point is further illustrated in Fig. 4 where we can see that a query image containing a bicycle is erroneously matched to instances of other people riding bicycles by the ViT baseline (Fig. 4, lower) whilst our approach performs correct Re-ID (Fig. 4, upper). Furthermore, across a range of video person Re-ID related challenges illustrated within Fig. 5(a-e) we can see that our approach successfully deals with issues such as scaling, human pose variation and motion blur (Fig. 5, left) whilst the baseline ViT [25] is shown to fail more prominently in these cases (Fig. 5, right).

Figure 5: Comparison of top-6 retrieval results on MARS [56] using our proposed method (VID-Trans-ReID) and ViT baseline (*green* = *correct Re-ID*, *red* = *incorrect Re-ID*).

6 Conclusion

In this paper, we propose an enhanced video transformer architecture for video person Re-ID using our novel Temporal Clip Shift and Shuffled (TCSS) and Video Patch Part Feature (VPPF) modules, in combination with both camera (view) embedding and current best practice approaches in video Re-ID. Quantitatively our approach outperforms all recent prior work in the field on established video Re-ID benchmarks (Rank-1 (mAP) – MARS: 96.36%(90.25%), iLIDS: 94.67%, PRID: 96.63) whilst qualitatively we illustrate enhanced attention maps with superior focus given to on-person visual features. Furthermore, we show the impact of adding our TCSS and VPPF modules to alternative transformer architectures (Swin + Focal) where we additionally show enhanced performance against the baseline.

References

[1] A. Aich, M. Zheng, S. Karanam, T. Chen, A.K. Roy-Chowdhury, and Z. Wu. Spatio-temporal representation factorization for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 152–162, 2021.

[2] A. Alsehaim and T.P. Breckon. Not 3D Re-ID: Simple single stream 2D convolution for robust video re-identification. In *Proc. Int. Conf. Pattern Recognition*, pages 5190–5197. IEEE, September 2020.

[3] A. Alsehaim and T.P. Breckon. Re-id-ar: Improved person re-identification in video via joint weakly supervised action recognition. In *Proc. British Machine Vision Conference*. BMVA, November 2021.

[4] S. Bai, B. Ma, H. Chang, R. Huang, S. Shan, and X. Chen. SANet: Statistic attention network for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[5] S. Bai, B. Ma, H. Chang, R. Huang, and X. Chen. Salient-to-broad transition for video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7339–7348, 2022.

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[7] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1169–1178, 2018.

[8] G. Chen, Y. Rao, J. Lu, and J. Zhou. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In *European Conference on Computer Vision*, pages 660–676. Springer, 2020.

[9] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8351–8361, 2019.

[10] D. Chung, K. Tahboub, and E.J Delp. A two stream siamese convolutional neural network for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1983–1991, 2017.

[11] A. Dehghan, S. Modiri Assari, and M. Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4091–4099, 2015.

[12] C. Eom, G. Lee, J. Lee, and B. Ham. Video-based person re-identification with spatial and temporal memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12036–12045, 2021.

[13] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1627–1645, 2009.

[14] J. Gao and R. Nevatia. Revisiting temporal modeling for video-based person reid. *arXiv preprint arXiv:1805.02104*, 2018.

[15] X. Gu, H. Chang, B. Ma, H. Zhang, and X. Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *European Conference on Computer Vision*, pages 228–243. Springer, 2020.

[16] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15013–15022, 2021.

[17] T. He, X. Jin, X. Shen, J. Huang, Z. Chen, and X. Hua. Dense interaction learning for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1490–1501, 2021.

[18] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[19] M. Hirzer, C. Beleznai, P.M Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian Conference on Image Analysis*, pages 91–102. 2011.

[20] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen. Vrstc: Occlusion-free video person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7183–7192, 2019.

[21] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen. Temporal complementary learning for video person re-identification. In *European Conference on Computer Vision*, pages 388–405. Springer, 2020.

[22] R. Hou, H. Chang, B. Ma, R. Huang, and S. Shan. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2014–2023, 2021.

[23] T. Isobe, D. Li, L. Tian, W. Chen, .Y Shan1, and W. Shengjin. Towards discriminative representation learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8526–8536, 2021.

[24] X. Jiang, Y. Qiao, J. Yan, Q. Li, W. Zheng, and D. Chen. Ssn3d: Self-separated network to align parts for 3d convolution in video person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1691–1699, 2021.

[25] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *Proc. Int. Conf. Learning Representations*, 2021.

[26] J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang. Global-local temporal representations for video person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3958–3967, 2019.

[27] J. Li, S. Zhang, and T. Huang. Multi-scale temporal cues learning for video person re-identification. *IEEE Transactions on Image Processing*, pages 4461–4473, 2020.

[28] M. Li, H. Xu, J. Wang, W. Li, and Y. Sun. Temporal aggregation with clip-level attention for video-based person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3376–3384, 2020.

[29] S. Li, S. Bak, P. Carr, and X. Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2018.

[30] Y. Li, X. Weng, Y. Xu, and K.M. Kitani. Visio-temporal attention for multi-camera multi-target association. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9834–9844, 2021.

[31] J. Liu, Z. Zha, W. Wu, K. Zheng, and Q. Sun. Spatial-temporal correlation and topology learning for person re-identification in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4370–4379, 2021.

[32] X. Liu, P. Zhang, C. Yu, H. Lu, and X. Yang. Watching you: Global-guided reciprocal learning for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13334–13343, 2021.

[33] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5790–5799, 2017.

[34] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[35] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, pages 2597–2609, 2019.

[36] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, and C. Zhang. Alignedreid++: Dynamically matching local information for person re-identification. *Pattern Recognition*, pages 53–61, 2019.

[37] N. McLaughlin, J.M Del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2016.

[38] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 542–551, 2019.

[39] P. Pathak, A.E. Eshratifar, and M. Gormish. Video person re-id: Fantastic techniques and where to find them (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13893–13894, 2020.

[40] L. Peng, Z. Chen, Z. Fu, P. Liang, and E. Cheng. Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs. *arXiv preprint arXiv:2203.04050*, 2022.

[41] Z. Qiu, T. Yao, and T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[42] A. Subramaniam, A. Nambiar, and A. Mittal. Co-segmentation inspired attention networks for video-based person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 562–572, 2019.

[43] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision*, pages 480–496, 2018.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[45] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[46] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

[47] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the International Conference on Multimedia*, pages 274–282, 2018.

[48] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6449–6458, 2020.

[49] W. Wang, E. Xie, X. Li, D.P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, pages 1–10, 2022.

[50] X. Wang and R. Zhao. Person re-identification: System design and evaluation overview. In *Person Re-Identification*, pages 351–370. Springer, 2014.

[51] Y. Wang, P. Zhang, S. Gao, X. Geng, H. Lu, and D. Wang. Pyramid spatial-temporal aggregation for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12026–12035, 2021.

[52] Z. Wang, L. He, X. Tu, J. Zhao, X. Gao, S. Shen, and J. Feng. Robust video-based person re-identification by hierarchical mining. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2021.

[53] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. 2016.

[54] D. Wu, M. Ye, G. Lin, X. Gao, and J. Shen. Person re-identification by context-aware part attention and multi-head collaborative learning. *IEEE Transactions on Information Forensics and Security*, pages 115–126, 2021.

[55] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4733–4742, 2017.

[56] Y. Yan, J. Qin, J. Chen, L. Liu, F. Zhu, Y. Tai, and L. Shao. Learning multi-granular hypergraphs for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2899–2908, 2020.

[57] J. Yang, W. Zheng, Q. Yang, Y. Chen, and Q. Tian. Spatial-temporal graph convolutional network for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3289–3299, 2020.

[58] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021.

[59] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing*, pages 2860–2871, 2019.

[60] Z. Yu, J. Pei, M. Zhu, J. Zhang, and J. Li. Multi-attribute adaptive aggregation transformer for vehicle re-identification. *Information Processing & Management*, page 102868, 2022.

[61] X. Zang, G. Li, and W. Gao. Multi-direction and multi-scale pyramid in transformer for video-based pedestrian retrieval. *IEEE Transactions on Industrial Informatics*, 2022.

[62] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 2018.

[63] Z. Zhang, C. Lan, W. Zeng, and Z. Chen. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10407–10416, 2020.

[64] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen. Relation-aware global attention for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3186–3195, 2020.

[65] Y. Zhao, X. Shen, Z. Jin, H. Lu, and X. Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4913–4922, 2019.

[66] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *Proc. European Conference on Computer Vision*, pages 868–884. Springer, 2016.

[67] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13001–13008, 2020.

[68] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3702–3712, 2019.

[69] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4747–4756, 2017.

[70] K. Zhu, H. Guo, Z. Liu, M. Tang, and J. Wang. Identity-guided human semantic parsing for person re-identification. In *European Conference on Computer Vision*, pages 346–363. Springer, 2020.

[71] K. Zhu, H. Guo, S. Zhang, Y. Wang, G. Huang, H. Qiao, J. Liu, J. Wang, and M. Tang. AAformer: Auto-aligned transformer for person re-identification. *arXiv preprint arXiv:2104.00921*, 2021.

[72] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.