scientific reports

Check for updates

OPEN

Channel-shuffled transformers for cross-modality person re-identification in video

Rangwan Kasantikul[1,3], Worapan Kusakunniran[1]✉, Qiang Wu[2] & Zhiyong Wang[3]

Effective implementation of person re-identification (Re-ID) across different modalities (such as daylight vs night-vision) is crucial for Surveillance applications. Information from multiple frames is essential for effective re-identification, where visual components from individual frames become less reliable. While transformers can enhance the temporal information extraction, the large number of channels required for effective feature encoding introduces scaling challenges. This could lead to overfitting and instability during training. Therefore, we proposed a novel Channel-Shuffled Temporal Transformer (CSTT) for processing multi-frame sequences in conjunction with a ResNet backbone to form Hybrid Channel-Shuffled Transformer Net (HCSTNET). Replacing fully connected layers in standard multi-head attention with ShuffleNet-like structures is important for integration of transformer attention with a ResNet backbone. Applying ShuffleNet-like structures reduces overfitting through parameter reduction with channel-grouping, and further improves learned attention using channel-shuffling. According to our tests with the SYSU-MM01 dataset in comparison against simple averaging of multiple frames, only the temporal transformer with channel-shuffling achieved a measurable improvement over the baseline. We have also investigated the optimal partitioning of feature maps therein.

Keywords Channel Shuffling, Transformer, Surveillance, Video-Based Person Re-identification, RGB-IR Re-ID

Person re-identification (Person Re-ID) is one of the major applications in surveillance. Sample-to-Sample variations of the same person have always been a challenge for successful implementation of Re-ID techniques. Footages are expected to have changes in viewing angles, in the amount of illumination, and in background compositions, all of which negatively affect the accuracy of the Re-ID system. More importantly, most modern cameras are capable of switching into infrared mode during a low-light situation. While this enabled the cameras to record footages in the dark, the resulting outputs would have radically different characteristics compared to the ones recorded in visible light. Any Re-ID models unprepared for such scenarios will struggle to make accurate identifications. Cross-modality techniques[1–5] handles such variances by imparting modality awareness as part of the techniques.

Beside modality awareness, Re-ID performance can be improved by using multiple frames as the input. The advantageous uses of multiple-frame or motion information have long been established in the field of Re-ID on single-modality scenarios[6,7]. Models utilizing multi-frame inputs can exploit temporal information that does not exist on a single input frame. In addition, combining multiple frames have also been used for gait recognition[8]. According to Nambiar et al.[9], gait information can be one of the strongest identifiers in the absence of high-quality appearance information. These are particularly important in cross-modality Re-ID, when appearance-based information such as colours and textures become even less reliable than single-modality Re-ID scenarios. This is due to the yet greater levels of sample-to-sample variances between images across modalities.

Multiple frames used for Re-ID models can be combined in many fashions[6]such as simple pooling, recurrent neural network (RNN), or include attention models[7,10]to focus on the essential parts of the input sequences. Among implementations of attention models, transformer models[11]have seen popularity in natural language processing and is regarded as the superior successor to RNN. This is due to the transformer's superior ability to capture long-range dependency. In addition, transformer uses a multi-head version of self-attention, allowing it to capture multiple dependencies without requiring deep stacks of attention blocks. Transformers have also

[1]Faculty of Information and Communication Technology, Mahidol University, 999 Phuttamonthon 4 Road, Salaya 73170, Nakhon Pathom, Thailand. [2]School of Electrical and Data Engineering, University of Technology Sydney, 15 Broadway, Ultimo 2007, New South Wales, Australia. [3]School of Computer Science, The University of Sydney, Camperdown 2006, New South Wales, Australia. ✉email: worapan.kun@mahidol.edu

seen uses in processing visual inputs[12]. These typical vision-based transformer implementations apply attention mechanisms on one image or frame at a time.

One of the potential shortcomings in transformer-based architecture is the quadratic complexity of computation with respect to token length. This led to the development of SWIN[13] as the compromise, where the tokens are confined within a sliding window. However, Wu et al.[14]demonstrated in convolutional transformer (CvT) that a model can benefit from both local bias of convolutional neural networks (CNN) and non-local attentions of transformers. Token length aside, the token size (i.e. number of token channels) can also pose a problem for transformer-based solutions. This is due to the similar quadratic relationship of parameter footprints with respect to the token size, owing to the underlying scaled dot product inside the transformer modules. For large token sizes, the parameters required can be large enough to cause overfitting and instability during training, as well as computational burden. To make this integration viable, CvT[14]used depthwise-separable projections to minimize parameters. However, depthwise separation implies no interaction between channels that is essential for maintaining high-quality representation of the output features[15]. The channel-shuffling idea from ShuffleNet also inspired token shuffling between windows in Shuffled Transformer[16], which restores global-attention capabilities in SWIN and improves its performance.

Likewise, we saw an opportunity to explore the use of channel-shuffling to address the parameter complexity caused by token size. With this in mind, we decided to combine the existing cross-modality approach with multi-frame inputs. The cross-modality models based on CNN serve as a good baseline to generate modality-agnostic features from visual footage. Multiple frames of inputs combined can provide the gait information, which further enhances the identifiability of the output features. The transformer is then used temporally to capture temporal attention and enhance the final feature. In this instance, the output frames from the CNN backbone network are treated the same as tokens for transformer. There, we utilize the channel-shuffling for tokens to manage the parameter footprint and achieve performance improvement over baseline. Following this idea, our work proclaims the contributions as follows:

- Channel-shuffling has been introduced for the transformer's attention module, resulting in a smaller parameter footprint and improved performance over standard transformer Implementations.
- This work demonstrated that parameter reduction by channel-grouping in transformer can address the performance degradation in application with large token sizes, and that channel-shuffling further enhances the quality of the attention learned from the now-reduced parameters.
- Channel-shuffled transformer architecture has been tested on a cross-modality Re-ID technique based on multi-frame inputs using CNN with temporal transformers on large-token sequences.
- Partitioning strategies of the feature maps have been investigated for optimum Re-ID performance on the SYSU-MM01 dataset.

Related work

Cross-modality Re-ID techniques

To handle the challenges that arise with the nigh-vision cameras is how to identify the person captured in the infrared mode, research efforts are made in the cross-modality re-identification. The cross-modality Re-ID problem[2]is the identification of people captured in infrared mode by comparing it to the previously registered identities in coloured mode. Therefore, the strong model for this task must be able to overcome the differences between the coloured inputs and infrared inputs to make an effective identification of the subject. The state-of-the art techniques for cross-modality Re-ID were developed in various configuration. The most straightforward modification in cross-modality Re-ID implementation is by making the loss-functions modality-aware. This encourages the model to make the closest match between same-person inputs across modalities rather than different-person within the same modality[3,17–19].

Cross-modality Re-ID techniques can be classified by how different modalities are handled in the model. Some techniques prefer[5,20,21]sharing parameters between two modalities and focus on modality-aware losses. Other techniques provide separate pipelines and parameters for different modalities. The approaches[1,17,22]belonging to the latter group tend to provide better performance potential, but also incurred modality dependency where modality labels are required during the inference/deployment stage. Nevertheless, the pipeline distillation[4,23] may also be used to alleviate the dependency problem during the inference. However, this approach also requires even more pipelines, which results in additional hardware burden during the training.

In addition to modality-aware losses, models can be trained to produce modality-agnostic features through adversarial learning. These come in the form of label-adversarial learning or generative adversarial network (GAN). In the case of GAN, the generator is used to create false images of another modality[24,25]. These false images can then be treated the same as the target modality by the feature extractor. While these techniques require minimal modifications to the feature extractor to be modality-aware, they require multiple steps of training. Furthermore, even though the Re-ID model itself operates on the same modality, it remains reliant on modality conversion by the generator. Therefore, the false modality generator is required for subsequent inferences, which imposes additional hardware burden and modality dependency.

On the other hand, the techniques based label-adversarial learning[20,26] do not require a separate generator from the feature extractor. The feature extractor itself acts as the generator in adversarial learning, which will be trained to extract modality-agnostic features. A modality discriminator is used in training to judge whether the features created by the model are close enough to be considered indistinguishable. Compared to GAN, there is no need for extra false image generators and the models can therefore be made smaller. However, the training process remains somewhat complex because the generator and the modality discriminator cannot be updated in the same pass.

Some techniques argued that the modality-gap between visible and infrared is simply too large to be properly trained on their own. Instead, these techniques employed intermediate-modality input to facilitate the training process. The "intermediate" modalities can be greyscale image derived from colour input[21,27], edge information[22], or patches of two modalities stitched together[28]. However, this comes at the cost of larger batch size required and possible extra pipeline during training.

Considering the advantages and disadvantages of these techniques, we decided to select the parameter-shared approach as the baseline. The performance disadvantages are compensated by not needing modality labels during inferences. Furthermore, it is still possible to improve the Re-ID performance by introducing an intermediate modality. Conventionally, this requires an additional branch in the model. With fully shared parameters in Homogeneous Augmented Tri-Modal Learning (HAT) technique[21], however, this extra modality requires no modification to the model architecture. This simplicity makes it ideal for our study.

Multi-frame processing through attention models

Gao et al.[6]has proposed a technique based on ResNet50 for identifying person from video footages. The ResNet was used to convert each video frame into a feature map. The work has offered the comparison between different approaches in exploiting temporal data from video sequences (e.g., RNN, temporal pooling, temporal attention), and concluded that temporal attention (which assigns weights based on visual attention on each frames to emphasize most important frames) is the most suitable approach for person identification from video footages. Interestingly, the work also shown that the simplest method of temporal pooling can outperform both gated recurrent unit (GRU) and long-short term memory (LSTM). The temporal pooling method was only narrowly defeated by temporal attention. Karpathy et al., who presented a similar technique for video classification (averaging class logits across the clip)[29] earlier, noted that the average of class logits from the clip formed a more robust prediction compared to individual predictions, even without specialized technique for motion capture.

The successor to RNN architectures is transformers, which have their roots from natural language processing (NLP). One of the major components in the transformers model is multi-headed scaled dot product attention (MHA). This allowed transformers to outperform RNN architectures in large-text translation tasks[11]. The success of transformers in natural language processing inspired adoptions in different subject areas, such as language models[30,31], text summarization[32]and time-series predictions[33]. Vision transformers (ViT)[12]is one of the adaptations of transformers model in computer vision. Patches of images are embedded and treated as sequence tokens, similarly to words in sentences (in a body of text). In comparison to convolutional neural networks (CNNs), ViT enjoyed non-locality and suffered less inductive spatial biases. This enabled transformers to perform well on larger datasets. However, recent CNN designs[34]also disputed such benefits, claiming that such spatial biases are necessary sacrifices to keep the model at a reasonable size and also necessary for the models to perform well on smaller datasets. Later on, derivatives of ViT such as SWIN[13] were developed to reach such a compromise.

An architecture can be developed using the components from both CNN and transformer to combine the strengths of these two designs, as demonstrated by CvT[14]. The hybrid approach of transformer and CNN combination is also emerging in Re-ID domain[35]. These architectures, however, only designed around a single input image. These developments became an inspiration for us to explore other possible approaches using transformer architecture. Instead of spatial locations, CNN output features of individual images across the sequences became tokens for transformers. This hybrid approach leaves the feature map extraction of image sequences to the CNN backbone, producing a string of feature maps. The individual feature maps are fed to the transformer layer to determine attention like the tokens in natural language processing. In this approach, both constituent architectures are assigned respectively to their familiar tasks to perform optimally.

Feature partitioning

Yao et al.[36] demonstrated the importance of part-based learning for Re-ID applications. In the absence of feature partitioning, the features would have a tendency to focus on the midsection of the subjects. This behaviour may be caused by the model using clothings as the discriminative factor. The non-involvement of other body parts (i.e. legs, arms, and head) in the feature representation negatively impacts the generalizability of the trained model. According to Sun et al.[37], the preceding Re-ID techniques with part-based learning can be categorized into two major groups. The first group are those relying on external pose information. Since explicit pose information is not always readily available from the input images, these techniques would require pose annotation from human inputs, or rely on a pre-trained pose-estimator for semantic segmentation[38,39]. This approach eliminates the need for human input or external model. However, the noise from segmentation remained an issue for the approaches of this category.

Sun et al.[37]also proposed a feature partitioning framework named Part-based Convolutional Baseline (PCB). The work suggested that the feature map from the CNN backbone can be divided along the height (rows) of the map. The partitions can then be used simultaneously and individually for representative learning, which enforces consistency between parts. The partitioning can either be made uniformly or undergo further refinement (at the sacrifice of end-to-end training). The work discovered that for the input with an aspect ratio of 3:1, the optimum number of height-wise partitions is 6. The excessive amount of partitioning may lead to training collapse due to some individual parts being too similar to one another. The number of partitioning would vary for inputs of different aspect ratios. Some adaptations retained the 6-partition on inputs with 2:1 aspect ratio[40,41]. Others would adapt a different partitioning number for their use cases[1,42]. All in all, while partitioning can help improve performance, the number of optimal partitioning is not universally agreed upon and may require trials-and-errors for individual techniques. Furthermore, 2D-partitioning (on both rows and columns) for the final feature remains a venue to be explored.

Complexity reduction with grouping and shuffling

Attempts to reduce complexity using grouping is dated back to AlexNet[43]as an approach to fit the model across two inexpensive GPUs (NVIDIA GTX 580). The CNN layers are split into two groups – one for each GPU. The layers across the two groups run largely independently, with crosstalks only on a few layers. Apart from the benefits of smaller overall footprints and parallelizability, grouping and performing separate convolution was also later found to alleviate overfitting[44]. This grouping technique has also become a basis of other CNN architectures, such as ResNeXt[45].

Although grouped convolution can achieve reduction in complexity, the connections among feature channels have been sacrificed due to isolation between groups. This lack of channel-interaction between groups could weaken the overall feature representation. To remedy this, ShuffleNet[15] introduced a channel-shuffling module to rearrange the feature channels such that the channels from different groups will interact in the subsequent grouped convolution and strengthen the representation. This significantly improved the prediction performance than grouped convolution only. The model also performed noticeably better compared to the similarly sized MobileNet.

Similar shuffling ideas have also seen adoption in vision transformers (ViT). Despite performing well in classification problems, application of ViT in other areas is limited due to scaling challenges[13]. This scaling issue is caused by the quadratic complexity with respect to token length. The early attempt in addressing this challenge was the SWIN architecture, which employed hierarchical windowing to limit the number of tokens to be processed at one time. While this can be considered mimicking a CNN, this can also be considered the equivalent to grouping in ShuffleNet without shuffling, where computational complexity is reduced at the sacrifice of global attention. This trade-off has similarly been rectified in Shuffle Transformer[16], where tokens across windows are rearranged. This leaves one other aspect for us to explore – the scaling problems with large token sizes, which could possibly be mitigated with channel-shuffling.

Proposed methods

Overview

The feasibility of combining the CNNs with transformers culminated in the design of Hybrid Channel-Shuffled Transformer Net (HCSTNET) as shown in Fig. 1, containing both CNN backbone and transformer attention layers. The input being sequences of images of length *f*. The model first utilizes a CNN backbone to learn feature

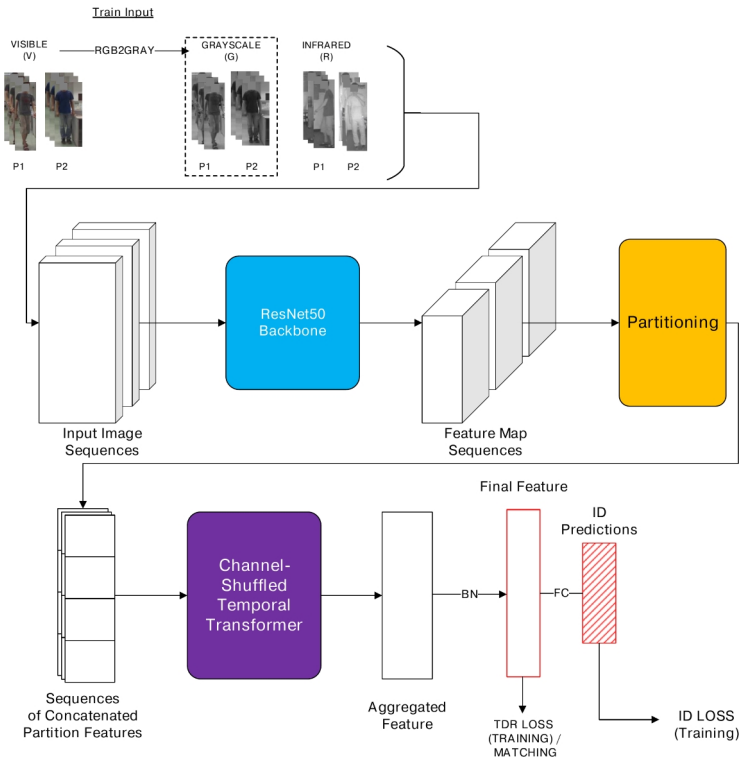

Fig. 1. The overall framework of HCSTNET architecture.The overall framework of HCSTNET architecture. The model contains The feature encoder consisting of ResNet50 Backbone. The individual feature maps are then partitioned (section "Feature partitioning") and reduced to vectors of few spatial regions, which are then concatenated and used as tokens for the temporal transformer. Multi-frame processor is based on Channel-Shuffled Temporal Transformer (Fig. 3, section "Channel-shuffled temporal transformer") responsible for assigning weights to tokens – the part-based representation of individual frames in the sequences. Loss module (section "Loss function") consisting of batch-normalization (BN) with ranking loss (TDR) attached to the output, and with the same output attached to a classifier layer to take the ID loss (cross-entropy).

encoding from input images. There are several network architectures available for this purpose, such as ResNet or ViT. Among these choices, we use the backbone based on the commonly used ResNet50[6,7,21,23].

Afterwards, The feature maps from the backbone would undergo partitioning to preserve coarse-grained spatial information before being pooled and concatenated. The partitioned-concatenated features in the sequences form the feature tokens for the temporal transformer, which provides focus on key frames across the sequences and produces the refined, temporally aggregated features. The features then go through a batch-normalization layer. The output at this stage is used in evaluation. During training, the features from the backbone are then further transformed by the fully connected layers into class predictions of training subjects.

The model is to be trained end-to-end with mini-batches of training samples. The model is optimized with ranking loss, and is also connected to a classifier for optimization with ID loss (softmax). To optimize the training for cross-modality, we use the modality-aware tri-directional ranking loss (TDR) from HAT technique[21] as the ranking loss. At this stage, the input to the model are mini-batches of image sequences of both infrared and visible modalities from subjects assigned to the training set. To facilitate the cross-modality training and satisfy the requirements of TDR loss, a greyscale modality derived from visible inputs is supplied along with the training batch. The greyscale modality is used during training only.

Once the training is complete, the model can then undergo evaluation or be used for re-identification. At this stage, the model is used to generate features using the images of the known subjects for later identifications. These generated features are assigned labels according to the identities of the subjects and stored in the database as the gallery. Once the identification is needed for a query of an unidentified subject image, the same model is used to generate the query features. For the cross-modality scenario, both the gallery and query will originate from different modalities. These features are then compared with the gallery and then return the closest matches based on distance metrics – typically Euclidean[6]or Cosine Distance[46]. The latter is chosen as the metric for the ranking loss and during inference.

Feature partitioning

The initial output from the feature encoder is a set of activation maps, which still contain significant spatial dimensions. To avoid overfitting, the dimensions must further be reduced before use in the downstream task. One popular approach is to simply discard the dimensions through global average pooling (GAP)[6]. However, outright removing the spatial dimensions can undermine the discriminative properties of the final feature vector. In particular, images of a person can be divided into upper (head, arms, and top-wear) and lower (feet, legs, and bottom-wear) parts, which independently contribute towards identification.

To achieve a compromise between avoiding overfitting and preserving individual part-features, each feature map is partitioned and then pooled to a specific output dimension. The details of the process are shown in Fig. 2 based on AdaptiveAveragePooling, which in turn is based on average pooling with window size and stride based on the sizes of the input and the output. The output is flattened and concatenated channel-wise and is used as a token for the temporal transformer.

Channel-shuffled temporal transformer

As discussed earlier in Section "Multi-frame processing through attention models", there are several approaches for combining multi-frame features to represent an entire sequence, from simple averaging to attention models (which assign weights to individual frames before averaging). Following Fig. 1, We opted to use temporal transformer to provide the weights for frames in the sequences, which are treated as tokens in the transformer models. After the transformer modules, the weighted frame features go through temporal pooling, which averages the features across the frames and produces a single feature representing the entire sequence. This

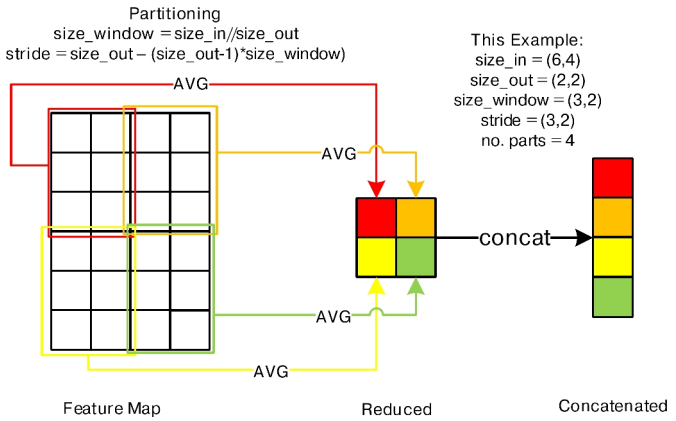

Fig. 2. The partitioning of a feature map based on AdaptiveAveragePooling to a specific spatial output dimension. This AdaptiveAveragePooling dynamically assigns the window size ($\text{size}_{\text{window}}$) and stride ($\text{stride}$) based on the input ($\text{size}_{\text{in}}$) and output ($\text{size}_{\text{out}}$) size. For the partitioning of 2D feature maps, the number of output rows and columns is to be supplied and calculated individually, which can be expressed as tuples of $(\text{rows}, \text{columns})$. The output is then flattened and concatenated along the channels.

whole-sequence feature is subsequently batch-normalized and then connected to loss functions during training. It is also used for re-identification during the testing/evaluation stage.

Internally, transformers rely on multi-head attention based on scaled dot product. The oncoming feature input is projected into query (Q), key (V), and value (V) and then multiplied to each other as seen in Fig. 3a. The reliance on scaled dot product complicates the integration of transformers with CNN due to the amount of new parameters required to accommodate the number of the output channels from the CNN. The naïve implementation of projection layer (and by extension, multi-head attention module) is performed using a regular fully connected layer. For an input feature of d channels, scaled dot product attention requires $3d^2$ parameters to project the input into Q, K, and V. As shown in Table 1, applying it on a ResNet50 output of 2048 channels would require an additional 12 million parameters for such projection per layer. Furthermore, the quadratic relationship between the number of channels and the total number of parameters required would rapidly increase as the input channels becomes larger. Having this many parameters in one layer would affect the

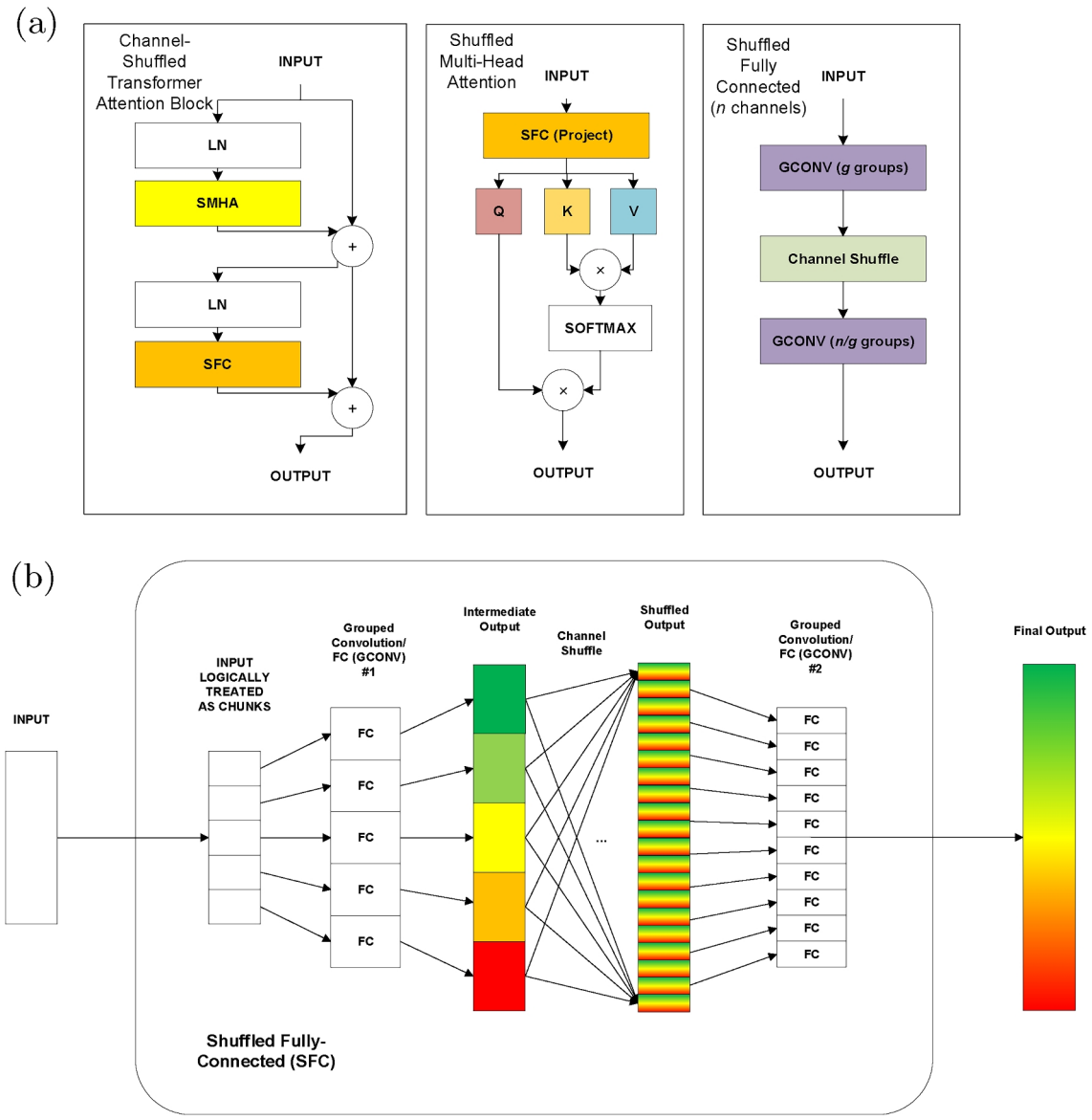

Fig. 3. The Channel-Shuffled Transformer Attention Block (CSTB) and the Shuffled Multi-Head Attention (SMHA) module used in HCSTNET as Channel-Shuffled Temporal Transformer (CSTT). Both modules use a Shuffled Fully Connected (SFC) module consisting of a channel shuffle layer sandwiched between two grouped convolutions. (**a**) The Channel-Shuffled Transformer Attention Block (CSTB), the Shuffled Multi-Head Attention (SMHA), and the Shuffled Fully Connected (SFC) module. (**b**) The detailed view of the Shuffled Fully Connected (SFC) module. Grouped convolution (GCONV) layers contain groups of fully connected layers (or the functionally equivalent pointwise convolution) operating on chunks of the input. To ensure interactions between chunks, a channel shuffle layer is used to reorder the chunks before forwarding to the second GCONV. The number of groups in the second GCONV is reciprocal to the number of groups in the first GCONV. Note: some connections were omitted for brevity.

No. Parts	No. Channels	No. Params	No. Params with shuffling		
		w/o shuffling	Groups = 8	Groups = 16	Groups = 32
1	2048	**12.6M**	1.7M	1.1M	983.0 K
2	4096	**50.3M**	6.6M	3.7M	2.8M
4	8192	**201.3M**	25.8M	13.8M	8.7M
6	12288	**453.0M**	57.5M	30.1M	17.7M
8	16384	**805.3M**	101.8M	52.7M	29.9M

Table 1. Comparison of parameters required in QKV projection layer for a given number of channels between standard and channel-shuffled implementation. Due to quadratic complexity between the number of channels and the number of parameters needed, the total parameters required quickly becomes unmanageable as the number of channels increases (as partitioning number – No. Parts increases). The use of channel-shuffling can dramatically reduce the parameter number compared to standard implementation.

trainability of the model, potentially causing overfitting and instability in training. In addition, implementations of transformers usually involve stacking multiple attention blocks, which further exacerbate the parameter number issue.

One way to address this parameter scaling problem is to replace the fully connected (FC) layers in transformer attention block with grouped convolution[43] (1D with kernel size of 1). However, while grouped convolution can reduce the parameter numbers, this approach leaves features from different groups isolated from interaction, which limits the quality of the learned projection. Shuffled Fully Connected (SFC) layers (Fig. 3b) based on the idea from ShuffleNet[15] were introduced for this purpose. The SFC layer contains a channel shuffle layer sandwiched between two grouped convolutions. With this idea, we can reduce the number of parameters from $3d^2$ to $3d^2/g + 9dg$ when using g groups in convolutions. Based on the same example with 2048-dimension inputs, we would require only 1 million parameters instead of 12 millions when 16-group convolutions are used. This results in Channel-Shuffled Transformer Attention Block (CSTB) for use in temporal transformer, becoming Channel-Shuffled Temporal Transformer (CSTT).

Loss function

During the training, the Re-ID models typically take cross-entropy and ranking losses[6] (Equation 1). The cross-entropy ($\mathcal{L}_{\text{CrossEnt}}$) loss, commonly utilized in classifier training, is also used to help model generate optimal feature representations for the inputs of each training subject. During the training, the batch-normalized output of the model is attached to the classification layer to produce a one-hot representation of the predicted labels. This classification layer is then discarded during inference. The ranking loss encourages the model to produce feature vectors that are similar within the same subjects, while also distinguishable among different subjects. The triplet losses ($\mathcal{L}_{\text{Tri}}$) are the popular choices among the ranking losses.

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{CrossEnt}} + \mathcal{L}_{\text{Tri}} \tag{1}$$

However, neither the ResNet50 backbone nor the triplet loss is specifically designed for cross-modality Matching. This can be rectified by modifying the training loss to be modality-aware. In addition, the network architecture can also be modified to handle different modalities separately, at the cost of modality dependency. Homogeneous Augmented Tri-Modal Learning (HAT) technique[21] demonstrated that much of the cross-modality performance can be gained by modifying only the training loss. This technique is chosen as the baseline in our implementation. The most defining features of the HAT technique are the introduction of greyscale input derived from the visible images and the use of tri-directional ranking loss (TDR, Equation 2). The derived greyscale images share the lack of colour with the infrared images, but keep the structures of the visible images. Therefore, the introduced greyscale modality bridges the modality gap between the visible and the infrared image. Based on this extra input, the TDR loss pulls the positive pairs (similarly to triplet loss) across different modalities together and pushes away the negative pairs.

$$\mathcal{L}_{\text{TDR}} = \mathcal{L}_{v,r,g} + \mathcal{L}_{r,g,v} + \mathcal{L}_{g,v,r} \tag{2}$$

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{TDR}} + \mathcal{L}_{\text{XE\_LS}} + \mathcal{L}_{\text{IFR}} \tag{3}$$

From the Equation 2, the terms $\mathcal{L}_{v,r,g}, \mathcal{L}_{r,g,v}, \mathcal{L}_{g,v,r}$ are the modality-constrained batch-hard triplet losses. For $\mathcal{L}_{v,r,g}$, the visible image (v) is the anchor, the infrared image (r) is the positive pair (i.e. sharing the same identity) to the anchor, and the greyscale image (g) is the negative pair (i.e. having different identity) to the anchor respectively. The other terms $\mathcal{L}_{r,g,v}, \mathcal{L}_{g,v,r}$ would have the modalities r, g, v taking on the different roles as the anchor and the positive/negative pairs to the anchor. The modality constraints on each term force the samples of different modalities to interact with one another.

The TDR loss is then combined with identity loss and other regularization functions. The $\mathcal{L}_{\text{TDR}}$ replaces $\mathcal{L}_{\text{Tri}}$ as the ranking loss from Equation 1. The ID loss $\mathcal{L}_{\text{XE\_LS}}$ used is a version of cross-entropy loss ($\mathcal{L}_{\text{CrossEnt}}$) with label smoothing[46]. The label smoothing (ϵ) adds noises on the ground truth and mitigates overfitting. As proposed in STA technique[7], the Inter-Frame Regularization ($\mathcal{L}_{\text{IFR}}$) was also used to enforce consistency under

the assumption that the frames of the same sequence should have similar activations. The $\mathcal{L}_{\mathrm{IFR}}$ is computed based on Frobenius-norm between a random pair of frames in the sequence.

Experiment

Datasets

The proposed techniques were tested on the SYSU-MM01 dataset[2], which provided both cross-view and cross-modality challenges. The dataset contains 491 identities, 296 for training, 99 for validation, and 96 for testing. Subjects are recorded by 6 cameras (4 in visible mode, 2 in infrared mode) in different locations. The footages taken under infrared mode are used as a *query* to match with gallery footages taken under visible mode. The dataset contains two main settings that can be set to alter the premises of the problem. The first setting is all-search vs. indoor-search mode. The indoor-mode only permits footages taken from indoors to be used in the evaluation. The all-search mode includes outdoor footages in the testing, which contains more complex background and induces a more profound changes in sample-to-sample backgrounds. The other setting is single-shot vs multi-shot mode. The single-shot mode only samples one gallery feature per subject per camera, whereas the multi-shot mode samples multiple (up to ten) gallery features per subject per camera.

Implementation

The model was trained on the SYSU-MM01 dataset with Adams optimizer. Learning rate was first set with 10 epochs of linear warm-up to 10^{-4} followed by 4 cycles of Cosine Annealing (starts at 10^{-4}, then decayed to 0 over 100 epochs for each cycle) for the total of 400 epochs. For each iteration in the epochs, 5 random subjects from the training set were chosen, with 4 instances for each (batch size totalling 20). Each instance contained a contiguous sequence of a randomly selected starting point with a set length of k from each modality. With k set to 1, the model emulates the Re-ID on still images. During the training, all images were pre-processed with resizing (288×144 pixels) and colour normalization. In addition, random cropping, random deletion[7], and random horizontal flipping were used as augmentation techniques for training. The backbone ResNet was initialized using the weights from ImageNet. The overall loss function could be expressed as the Equation 3. The label smoothing parameter ϵ was set to 0.1. The ranking loss L_{TDR} for the feature was the TDR loss with margin set to 0.3. The number of heads and groups in CSTB was set to 16. During the experimentation, there were up to 3 Shuffled Transformer Attention Blocks stacked back-to-back. The model with 0 (none) of these blocks was treated as the placebo model for reference purposes. The model was evaluated for mean average precision (mAP) and top-k accuracy on both all-search and indoor-search for every 10 epochs of training. Each evaluation contained 10 individual randomized tests, from which mAP and top-k accuracy were averaged and recorded. The results for each evaluation were logged, with mAP and top-k accuracy from the epoch with highest mAP based on all-search being used in reporting of the result.

Results and discussion

We conducted the experimentation with different model setups. We first studied the optimal partitioning strategy, then the inclusion of temporal transformer, and finally the different groupings of Channel-Shuffled Temporal Transformer (CSTT). The reference model was configured with no attention block and only receives single frame (0x NTAB, 1 F). This represents the baseline Re-ID technique that is only taking modality-aware losses. The reference model was also tested with 8-Frame sequences (0x NTAB, 8 F). This allowed comparison of performance improvements that can be gained simply by using multiple input frames. The next configuration was the inclusion of transformer attention blocks with few variations. Naïve Transformer Attention Block (NTAB) was the baseline transformer block implementation without grouped convolutions or channel shuffling. Channel-Grouped Attention Block (CGTB) had the FC layers in the block replaced with grouped convolution, but without shuffling. Finally, Channel-Shuffled Transformer Attention Block (CSTB) contained both grouped convolution and shuffling, which we further examined by varying the number of blocks stacked to each other (1x-3x CSTB). In addition, a version using gated recurrent unit (GRU) had also been used for comparison to represent the recurrent designs of temporal model. Afterwards, we compared the performance level of our technique with existing ones using the same dataset.

Partitioning study

The first study to be discussed is partitioning strategy. This was performed to examine the effects of different partitioning schemes on the model performance. In addition, the outcome of this study would also be used to find the optimum baseline for further studies with temporal transformers. Table 2 shows the comparison of the baseline model performance under different partitioning strategies and sequence lengths. The partitioning is expressed using the number of output rows and columns (before flattening) as a tuple (rows, columns). The model was tested using either sequences of 1-Frame and 8-Frame inputs. The table reported mAP and top-k accuracy numbers on both *All-Search* and *Indoor-Only* search. Among these metrics, the mAP and top-1 accuracy (where the model makes exact match) on *All-Search* mode are given the most weight in consideration.

According to the results from Table 2, partitioning along the rows improved the performance up to a certain number of partitions, followed by a decline in both 1-Frame and 8-Frame settings when partitioning became excessive. Notably, the 1-Frame and 8-Frame setting performed best at different division strategy, with 1-Frame favouring (2,1) division and 8-Frame preferring the finer (4,1) partitioning. Partitioning the feature map along the columns (e.g. (2,2), (3,2), (4,2) partitioning) did not appear to positively affect the performance. The outcome of this study echoes that Sun et al.[37], but also indicated that optimal partitioning is dependent on input aspect ratios and sequence length of the inputs to be aggregated. Furthermore, the result also suggested that the lateral spatial information in the feature maps were not different enough to improve the discriminative strength of the features.

Partitions	All-Search					Indoor-Only				
	mAP (%)	top-1 (%)	top-5 (%)	top-10 (%)	top-20 (%)	mAP (%)	top-1 (%)	top-5 (%)	top-10 (%)	top-20 (%)
1-Frame Results										
(1,1)	45.03	45.10	74.53	85.25	94.07	62.06	53.01	84.17	92.56	97.49
(2,1)	47.28	47.11	77.33	87.36	95.08	64.68	55.93	87.14	95.18	98.45
(4,1)	42.07	43.02	73.44	85.05	93.57	58.69	49.59	81.67	91.28	97.35
(6,1)	39.95	40.45	71.26	84.64	93.91	57.40	48.35	81.62	91.25	97.66
(8,1)	36.18	36.65	67.03	79.54	90.28	52.01	42.73	75.85	87.26	95.54
(2,2)	42.38	43.22	74.05	85.24	94.11	59.04	50.00	82.76	92.57	97.94
(3,2)	39.75	40.74	72.42	84.54	93.78	56.13	47.02	80.28	90.35	97.44
(4,2)	37.12	38.38	68.68	81.08	91.45	53.55	44.26	77.51	88.31	96.25
8-Frame Results										
(1,1)	77.54	79.09	94.27	97.36	99.21	88.32	85.25	98.38	99.35	100.00
(2,1)	78.87	80.50	95.06	97.67	99.20	88.46	85.97	96.35	98.86	100.00
(4,1)	79.57	80.71	95.67	97.66	98.83	88.41	85.48	96.72	98.17	99.92
(6,1)	78.00	79.43	95.54	97.57	99.05	87.80	84.23	97.11	99.06	99.93
(8,1)	76.54	78.26	94.42	97.30	98.79	84.71	81.34	94.99	97.74	99.67
(2,2)	77.37	79.36	95.67	97.99	98.88	86.98	85.29	95.45	97.63	99.75
(3,2)	75.68	77.00	94.42	97.61	99.54	85.77	81.53	96.85	98.58	99.41
(4,2)	76.58	77.51	95.46	97.89	99.30	85.21	82.31	95.53	98.74	99.54

Table 2. The baseline model results with various partitioning schemes on 1-Frame and 8-Frame sequences. The partitions are expressed in (rows, columns) based on the specified output dimension of AdaptiveAveragePooling before the flattening operation. The performance is expressed in mean average precision (mAP) and top-k (e.g. top-1, top-5, etc.) accuracy in *All-Search* (all cameras used) and *Indoor-Only* (only indoor cameras used).

Models	All-Search					Indoor-Only				
	mAP (%)	top-1 (%)	top-5 (%)	top-10 (%)	top-20 (%)	mAP (%)	top-1 (%)	top-5 (%)	top-10 (%)	top-20 (%)
0x NTAB (8 F)	79.57	80.71	95.67	97.66	98.83	88.41	85.48	96.72	98.17	99.92
1x GRU (8 F)	64.61	65.70	90.18	95.93	98.87	80.19	75.53	94.37	97.67	99.36
1x NTAB (8 F)	53.05	49.69	82.58	91.28	96.57	67.52	58.49	88.39	95.48	99.53
1x CGTB (8 F)	66.77	66.28	87.84	93.44	97.08	79.66	73.49	92.69	96.25	98.58
1x CSTB (8 F)	80.60	82.39	96.73	98.03	98.77	89.19	85.96	97.10	98.41	99.99
2x CSTB (8 F)	79.90	81.42	95.29	97.45	98.50	88.59	86.43	96.08	97.67	99.84
3x CSTB (8 F)	80.60	82.42	95.98	98.09	98.65	89.00	86.48	96.47	98.49	100.00

Table 3. The results of temporal transformers using sequences of 8 images (8 F). The models with varying blocks of Non-Shuffled Transformer Attention Block (NTAB) or Channel-Shuffled Transformer Attention Block (CSTB) stacked after the output of CNN backbone. In addition, a model using Gated Recurrent Unit (GRU) has also been used for comparison. The performance is expressed in mean average precision (mAP) and top-k (e.g. top-1, top-5, etc.) accuracy in *All-Search* (all cameras used) and *Indoor-Only* (only indoor cameras used).

Grouping and shuffling study

Following the partitioning study in the previous section, we moved to examine the effects of introducing transformer model temporally to enhance the feature aggregation. Furthermore, the effects of channel grouping and shuffling in transformer implementation were also studied. Here, variants of the transformer architecture were examined; standard implementation, variant with channel grouping, and the variant with grouping and shuffling (channel-shuffled). Later on, we performed the in-depth study on multi-layer transformer blocks and different grouping schemes to further explore the performance potential with temporal transformer. In this study, the best-performing 8-Frame model with (4,1) feature partitioning was chosen as the basis for this experimentation.

Table 3 shows the effects of using temporal transformer under various configurations. Similar to the previous study, the top-k accuracy and mAP are compared in this table. The naïve implementation of multi-head attention resulted in a sharp decline of performance (1x NTAB) over one without the multi-head attention (0x NTAB). This is caused by the extreme amount of additional parameters required in the projection layer of MHA, resulting in overfitting and instability in training. The situation somewhat improved using grouped convolution in the transformer block (1x CGTB), reaching the level of (1x) GRU baseline. This suggests that parameter reduction

Heads/Groups	All-Search					Indoor-Only				
	mAP (%)	top-1 (%)	top-5 (%)	top-10 (%)	top-20 (%)	mAP (%)	top-1 (%)	top-5 (%)	top-10 (%)	top-20 (%)
8	78.92	80.33	95.01	97.54	98.70	88.00	85.02	96.74	98.75	99.96
16	80.60	82.39	96.73	98.03	98.77	89.19	85.96	97.10	98.41	99.99
32	79.55	80.09	96.23	97.92	98.76	88.75	86.06	96.55	98.30	99.95
64	79.91	80.89	96.20	98.22	98.90	88.35	85.05	97.04	98.59	99.98

Table 4. Comparison of performance between different heads and groupings used in the temporal transformer. The model with one block of Channel-Shuffled Temporal Transformer (1x CSTB) is used for the experimentation. The performance is expressed in mean average precision (mAP) and top-k (e.g. top-1, top-5, etc.) accuracy in *All-Search* (all cameras used) and *Indoor-Only* (only indoor cameras used).

Models	All-Search					Indoor-Only				
	mAP (%)	top-1 (%)	top-5 (%)	top-10 (%)	top-20 (%)	mAP (%)	top-1 (%)	top-5 (%)	top-10 (%)	top-20 (%)
Base (0x NTAB, 1 F)	47.28	47.11	77.33	87.36	95.08	64.68	55.93	87.14	95.18	98.45
EAT-CMKD[17]	43.09	43.23	-	82.78	90.91	58.88	50.07	-	90.63	96.99
Yuan et al.[5]	53.45	55.61	-	90.67	96.04	68.58	62.45	-	94.25	98.41
HAT[21]	53.89	55.29	-	92.14	97.36	69.37	62.10	-	95.75	99.20
HMML_C[3]	60.44	63.63	-	-	-	-	-	-	-	-
SCC-MGL[19]	60.83	68.02	-	94.21	97.39	73.25	70.60	-	93.18	96.36
TSME[25]	61.21	64.23	-	95.19	98.73	71.53	64.80		96.92	99.31
cm-SSFT[23]	63.2	61.6	-	89.2	93.9	72.6	70.5	-	94.9	97.7
IDCRL[47]	63.92	62.35	-	92.79	97.68	73.55	71.64	-	95.92	99.25
PMCM[1]	71.16	75.54	-	97.49	99.30	84.33	81.52	-	98.99	99.71
Base (0x NTAB, **8 F**)	79.57	80.71	95.67	97.66	98.83	88.41	85.48	96.72	98.17	99.92
Ours (1x CSTB, 8 F)	80.60	82.39	96.73	98.03	98.77	89.19	85.96	97.10	98.41	99.99

Table 5. Model Performance tested on our testing conditions compared to the existing techniques. The performance is expressed in mean average precision (mAP) and top-k (e.g. top-1, top-5, etc.) accuracy in *All-Search* (all cameras used) and *Indoor-Only* (only indoor cameras used). Our base model with 1-Frame input (0x NTAB, 1 F) uses (2,1) feature partitioning while the models with 8-Frame input (8F) uses (4,1) feature partitioning based on best results from Table 2.

through grouped convolution indeed reduced overfitting compared to the naïve implementation (1x NTAB). However, both grouped convolution and GRU version were inferior to the temporal pooling (0x NTAB) version. These findings reflect that of Gao et al.[6] where most temporal models (e.g., 3D CNN, GRU, LSTM) failed to outperform the simple temporal pooling. In CGTB's case, this can also be explained by the lack of cross-talking between channel-groups, which limits the quality of the learned attention within the sub-channels.

Using the Shuffled version of MHA (1x CSTB) not only restored the performance of the model, but also offered a small advantage over the configuration without a multi-head attention layer. Stacking additional transformer blocks (2x, 3x CSTB) did not appear to further improve the overall performance. Further tests in different groupings in Table 4 revealed that the configuration with 16 shuffled groups and heads being the sweet spot for the temporal transformer deployed for this model. The results of this study indeed suggested that channel-shuffling in CSTB played a vital role in improving performance of temporal transformer in CSTT implementation.

Comparison study

To put the performance of our technique into perspective, we compared the performance numbers with the existing techniques tested on the same dataset. Table 5 shows the best model result on the SYSU-MM01 dataset compared to the state-of-the-art techniques. Our baseline implementation on singe-frame input (0x NTAB, 1 F) with (2,1) partitioning started at the lowest performance among the competing techniques. This level of performance is slightly below the HAT technique that we based on. The situation is different when operating on 8-frame input baseline (0x NTAB, 8 F) with (4,1) partitioning, where it outperformed all existing techniques in the table at 79.57% in mAP and 80.71% in top-1 accuracy. This performance increase is mostly due to stronger representation as a result of utilizing multiple frames as the inputs. The HCSTNET model with channel-shuffled multi-head attention (1x CSTB) and (4,1) partitioning further improved the performance, achieving 80.60% in mAP and 82.39% in top-1 accuracy under the all-search mode. This improvement is due to efficient and effective implementation of temporal transformer with CSTT/CSTB. Under the less demanding indoor-search mode, the model delivered 89.19% in mAP and 85.96% in rank-1 accuracy. Compared to the HAT technique by Ye et al. we used as the baseline for implementation, our model outperformed this technique by over 20% under all-search mode. Compared to the best techniques we are aware of so far, we maintained a 9.44% edge in mAP and a 6.85%

improvement (over PMCM) in rank-1 accuracy under all-search mode. This shows that combining features from multiple frames, particularly with optimum use of temporal transformer (as in CSTT) can help the model achieve better Re-ID performance than state-of-the-art techniques. It is to be noted, however, that the state-of-the-art techniques on the SYSU-MM01 dataset used in this comparison only operate on single-frame input and this comparison might not necessarily be a fair one.

Conclusion and future work

In this work, we demonstrated how channel-shuffling can improve parameter footprint and performance of transformer-based models where tokens contain large number of channels. In this scenario, channel-grouping reduced parameter size and overfitting. Channel-shuffling then allows information to flow between the groups and improves learned representation in each group. We applied this idea in cross-modality Re-ID as HCSTNET where multiple frames are used for matching, with transformer being used temporally to process feature-sequences from ResNet50 backbone. Under this test scenario, only the transformer set-up with channel-shuffling surpassed the baseline configuration in performance. This suggested that the benefits of channel-shuffling applies even for processing tokens in transformer. Although the improvement over the baseline appeared minor, these findings echoed that of Gao et al.[6], where only the best attention models offered advantage over simple averaging between frames by a small margin. Other attempts such as 3D CNNs and RNNs in this literature resulted in varying degrees of performance deterioration. Incidentally, we have also explored the partitioning strategy for Re-ID models. The findings suggested that the optimum partitioning for input aspect ratio of 2:1 indeed differed from 6 vertical partitions suggested in PCB[48], but also dependent on sequence length. It can be inferred that the average of feature maps can tolerate finer partitioning.

Nevertheless, this study also came with some limitations. We have only been able to evaluate our models on the SYSU-MM01 dataset. This can be seen as the limiting factor to our performance claims. There were plans to conduct a test on RegDB[49]dataset. Unfortunately, we were unable to obtain the permission from the original owner of this dataset. In addition, there are more approaches for implementing multi-head attention to be explored[50], which can offer additional advantage over standard MHA implementation. In future work, we plan to test channel-shuffled transformer on the wider ranges of tasks. This can also include applications on edge devices where efficiency under reduced complexity is essential. Furthermore, we plan to have a more extensive study in the optimal implementation of multi-head attention and transformer blocks for cross-modality Re-ID tasks.

Data availability

The SYSU-MM01 dataset can be obtained from third-party authors in[2] (https://github.com/wuancong/SYSU-MM01) under their permissions.

Code availability

The code is available on a GitHub repository: https://github.com/rangwank/channel-shuffled-transformer.

Received: 4 February 2025; Accepted: 24 April 2025
Published online: 29 April 2025

References

1. Qian, Z., Lin, Y. & Du, B. Visible-infrared person re-identification via patch-mixed cross-modality learning. *Pattern Recognition* **157**, 110873. https://doi.org/10.1016/j.patcog.2024.110873 (2025).
2. Wu, A., Zheng, W.-S., Gong, S. & Lai, J. RGB-IR person re-identification by cross-modality similarity preservation. *International Journal of Computer Vision* **128**(6), 1765–1785. https://doi.org/10.1007/s11263-019-01290-1 (2020).
3. Zhang, L. et al. Hybrid modality metric learning for visible-infrared person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications* **18**(1s), 1–15. https://doi.org/10.1145/3473341 (2022).
4. Zhang, D. et al. Dual mutual learning for cross-modality person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology* **32**(8), 5361–5373. https://doi.org/10.1109/TCSVT.2022.3144775 (2022).
5. Yuan, B., Chen, B., Tan, Z., Shao, X. & Bao, B.-K. Unbiased feature enhancement framework for cross-modality person re-identification. *Multimedia Systems* **28**(3), 749–759. https://doi.org/10.1007/s00530-021-00872-9 (2022).
6. Gao, J., Nevatia, R.: Revisiting temporal modeling for video-based person reid. CoRR **abs/1805.02104** (2018) arXiv:https://arxiv.org/abs/1805.02104
7. Fu, Y., Wang, X., Wei, Y. & Huang, T.S.: STA: spatial-temporal attention for large-scale video-based person re-identification. CoRR **abs/1811.04129** (2018) arXiv:https://arxiv.org/abs/1811.04129
8. Yao, L., Kusakunniran, W., Wu, Q. & Zhang, J. Gait recognition using a few gait frames. *PeerJ Computer Science* **7**, 382. https://doi.org/10.7717/peerj-cs.382 (2021).
9. Nambiar, A., Bernardino, A. & Nascimento, J.C.: Gait-based person re-identification: A survey. ACM Comput. Surv. **52**(2) (2019) https://doi.org/10.1145/3243043
10. Subramaniam, A., Nambiar, A. & Mittal, A.: Co-segmentation inspired attention networks for video-based person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 562–572 (2019). https://doi.org/10.1109/ICCV.2019.00065
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I.: Attention Is All You Need (2023). arXiv:https://arxiv.org/abs/1706.03762
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2021). arXiv:https://arxiv.org/abs/2010.11929
13. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. & Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows (2021). arXiv:https://arxiv.org/abs/2103.14030
14. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L. & Zhang, L.: CvT: Introducing Convolutions to Vision Transformers (2021). arXiv:https://arxiv.org/abs/2103.15808

15. Zhang, X., Zhou, X., Lin, M. & Sun, J.: ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices (2017). arXiv:https://arxiv.org/abs/1707.01083
16. Huang, Z., Ben, Y., Luo, G., Cheng, P., Yu, G. & Fu, B.: Shuffle Transformer: Rethinking Spatial Shuffle for Vision Transformer (2021). arXiv:https://arxiv.org/abs/2106.03650
17. Gao, G., Shao, H., Wu, F., Yang, M. & Yu, Y. Leaning compact and representative features for cross-modality person re-identification. *World Wide Web* **25**(4), 1649–1666. https://doi.org/10.1007/s11280-022-01014-5 (2022).
18. Huang, N., Liu, K., Liu, Y., Zhang, Q. & Han, J. Cross-modality person re-identification via multi-task learning. *Pattern Recognition* **128**, 108653. https://doi.org/10.1016/j.patcog.2022.108653 (2022).
19. Wang, Y., Xu, K., Chai, Y., Jiang, Y. & Qi, G. Semantic consistent feature construction and multi-granularity feature learning for visible-infrared person re-identification. *The Visual Computer* https://doi.org/10.1007/s00371-023-02923-w (2023).
20. Wei, Z., Yang, X., Wang, N. & Gao, X. Flexible body partition-based adversarial learning for visible infrared person re-identification. *IEEE Transactions on Neural Networks and Learning Systems* **33**(9), 4676–4687. https://doi.org/10.1109/TNNLS.2021.3059713 (2022).
21. Ye, M., Shen, J. & Shao, L. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE Transactions on Information Forensics and Security* **16**, 728–739. https://doi.org/10.1109/TIFS.2020.3001665 (2021).
22. Jiang, J. et al. Graph sampling-based multi-stream enhancement network for visible-infrared person re-identification. *Sensors* **23**(18), https://doi.org/10.3390/s23187948 (2023).
23. Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q. & Yu, N.: Cross-modality person re-identification with shared-specific feature transfer. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 13376–13386 . (2020) https://doi.org/10.1109/CVPR42600.2020.01339
24. Wu, A., Zheng, W.-S., Yu, H.-X., Gong, S. & Lai, J.: Rgb-infrared cross-modality person re-identification. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5390–5399 (2017). https://doi.org/10.1109/ICCV.2017.575
25. Liu, J., Wang, J., Huang, N., Zhang, Q. & Han, J. Revisiting modality-specific feature compensation for visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology* **32**(10), 7226–7240. https://doi.org/10.1109/TCSVT.2022.3168999 (2022).
26. Lin, X., Li, J., Ma, Z., Li, H., Li, S., Xu, K., Lu, G. & Zhang, D.: Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20941–20950 (2022). https://doi.org/10.1109/CVPR52688.2022.02030
27. Guo, J., Ye, Y., Du, H. & Hao, X. A triple-path global-local feature complementary network for visible-infrared person re-identification. *Signal, Image and Video Processing* **18**(1), 911–921. https://doi.org/10.1007/s11760-023-02789-4 (2024).
28. Pang, Z., Wang, C., Zhao, L., Liu, Y. & Sharma, G.: Cross-modality hierarchical clustering and refinement for unsupervised visible-infrared person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 1–1 (2023) https://doi.org/10.1109/TCSVT.2023.3310015
29. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. & Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725–1732 (2014). https://doi.org/10.1109/CVPR.2014.223
30. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019). arXiv:https://arxiv.org/abs/1810.04805
31. Zhong, H., Zhang, Q., Li, W., Lin, R. & Tang, Y. Kpllm-ste: Knowledge-enhanced and prompt-aware large language models for short-text expansion. *World Wide Web* **28**(1), 9. https://doi.org/10.1007/s11280-024-01322-y (2024).
32. yamini, p., daneshfar, f. & Ghorbani, A.a.: Kurdsm: Transformer-based model for kurdish abstractive text summarization with an annotated corpus. IRANIAN JOURNAL OF ELECTRICAL AND ELECTRONIC ENGINEERING **20**(4) (2024) https://doi.org/10.22068/IJEEE.20.4.3299http://ijeee.iust.ac.ir/article-1-3299-en.pdf
33. Yan, B., Zhao, G., Song, L., Yu, Y. & Dong, J. Precln: Pretrained-based contrastive learning network for vehicle trajectory prediction. *World Wide Web* **26**(4), 1853–1875. https://doi.org/10.1007/s11280-022-01121-3 (2023).
34. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T. & Xie, S.: A ConvNet for the 2020s (2022). arXiv:https://arxiv.org/abs/2201.03545
35. Zhao, Y. & Zhu, S. Occluded pedestrian re-identification via res-vit double-branch hybrid network. *Multimedia Systems* **30**(1), 5. https://doi.org/10.1007/s00530-023-01235-2 (2024).
36. Yao, H. et al. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing* **28**(6), 2860–2871. https://doi.org/10.1109/tip.2019.2891888 (2019).
37. Sun, Y., Zheng, L., Yang, Y., Tian, Q. & Wang, S.: Beyond Part Models: Person Retrieval with Refined Part Pooling (and a Strong Convolutional Baseline) (2018). arXiv:https://arxiv.org/abs/1711.09349
38. Zhao, L., Li, X., Zhuang, Y. & Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3239–3248 (2017). https://doi.org/10.1109/ICCV.2017.349
39. Li, Z. et al. Learning part-alignment feature for person re-identification with spatial-temporal-based re-ranking method. *World Wide Web* **23**(3), 1907–1923. https://doi.org/10.1007/s11280-019-00734-5 (2020).
40. Gamal, A., Shoukry, N. & Salem, M.A.-M.: Long-term person re-identification model with a strong feature extractor. In: 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS), pp. 74–79 (2021). https://doi.org/10.1109/ICICIS52592.2021.9694212
41. Li, H., Xu, L., Zhang, Y., Tao, D. & Yu, Z.: Adversarial Self-Attack Defense and Spatial-Temporal Relation Mining for Visible-Infrared Video Person Re-Identification (2023). arXiv:https://arxiv.org/abs/2307.03903
42. Alsehaim, A., Breckon, T.P.: Vid-trans-reid: Enhanced video transformers for person re-identification. In: 33rd British Machine Vision Conference 2022, BMVC 2022, November 21-24, 2022. BMVA Press, London, UK (2022). https://bmvc2022.mpi-inf.mpg.de/0342.pdf
43. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **60**(6), 84–90. https://doi.org/10.1145/3065386 (2017).
44. Ioannou, Y., Robertson, D., Cipolla, R. & Criminisi, A.: Deep roots: Improving cnn efficiency with hierarchical filter groups. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5977–5986. IEEE, Honolulu, HI, USA (2017). https://doi.org/10.1109/cvpr.2017.633 .
45. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K.: Aggregated Residual Transformations for Deep Neural Networks. arXiv (2016). https://doi.org/10.48550/ARXIV.1611.05431.
46. Luo, H., Gu, Y., Liao, X., Lai, S. & Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1487–1495 (2019). https://doi.org/10.1109/CVPRW.2019.00190
47. Zhu, X. et al. Information disentanglement based cross-modal representation learning for visible-infrared person re-identification. *Multimedia Tools and Applications* https://doi.org/10.1007/s11042-022-13669-3 (2022).
48. Sun, Y., Zheng, L., Yang, Y., Tian, Q. & Wang, S.: Beyond Part Models: Person Retrieval with Refined Part Pooling (and a Strong Convolutional Baseline). arXiv (2017). https://doi.org/10.48550/ARXIV.1711.09349 .
49. Nguyen, D.T., Hong, H.G., Kim, K.W. & Park, K.R.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**(3) (2017) https://doi.org/10.3390/s17030605

50. Kang, H., Yang, M.-H. & Ryu, J.: Interactive Multi-Head Self-Attention with Linear Complexity (2024). arXiv:https://arxiv.org/abs/2402.17507

Acknowledgements
This research project was partially supported by Faculty of Information and Communication Technology, Mahidol University.

Author contributions
Rangwan and Worapan wrote the main manuscript text. All authors reviewed the manuscript.

Funding
This research project is supported by National Research Council of Thailand (NRCT): (Contract No. N41A640225).

Declarations

Competing interests
The authors have no other competing interests than the aforementioned funding to declare that are relevant to the content of this article.

Compliance with ethical standards
This work is a secondary research on a publicly available dataset involving human subjects under permission from the dataset owner. The authors confirm that all procedures and protocols of human research are exempt from review board approval.

Additional information
Correspondence and requests for materials should be addressed to W.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.