Spatio-Temporal Representation Factorization
for Video-based Person Re-Identification

Abhishek Aich[*,2], Meng Zheng[1], Srikrishna Karanam[1], Terrence Chen[1],
Amit K. Roy-Chowdhury[2], and Ziyan Wu[1]

[1]United Imaging Intelligence, Cambridge MA, [2]University of California, Riverside CA

{aaich001@, amitrc@ece.}ucr.edu, {first.last}@uii-ai.com

Abstract

Despite much recent progress in video-based person re-identification (re-ID), the current state-of-the-art still suffers from common real-world challenges such as appearance similarity among various people, occlusions, and frame misalignment. To alleviate these problems, we propose Spatio-Temporal Representation Factorization (STRF), a flexible new computational unit that can be used in conjunction with most existing 3D convolutional neural network architectures for re-ID. The key innovations of STRF over prior work include explicit pathways for learning discriminative temporal and spatial features, with each component further factorized to capture complementary person-specific appearance and motion information. Specifically, temporal factorization comprises two branches, one each for static features (e.g., the color of clothes) that do not change much over time, and dynamic features (e.g., walking patterns) that change over time. Further, spatial factorization also comprises two branches to learn both global (coarse segments) as well as local (finer segments) appearance features, with the local features particularly useful in cases of occlusion or spatial misalignment. These two factorization operations taken together result in a modular architecture for our parameter-wise light STRF unit that can be plugged in between any two 3D convolutional layers, resulting in an end-to-end learning framework. We empirically show that STRF improves performance of various existing baseline architectures while demonstrating new state-of-the-art results using standard person re-ID evaluation protocols on three benchmarks.

1. Introduction

We consider the problem of video-based person re-IDentification (re-ID). Given a video tracklet of a person of interest, the task is to retrieve the closest match (which ideally should be the true match) among a gallery set of video

---

* This work was done during Abhishek Aich's internship with United Imaging Intelligence. Corresponding author: Srikrishna Karanam.
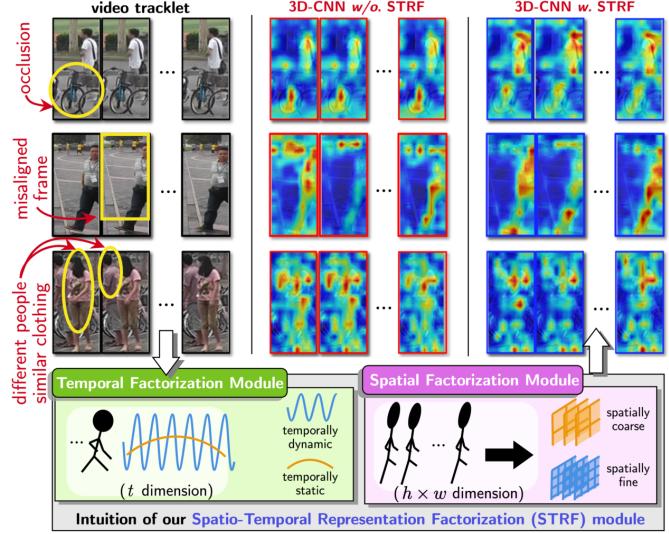

Figure 1: **Illustration of proposed concept and its efficacy.** We present the intuition behind our proposed Spatio-Temporal Representation Factorization (STRF) module designed to overcome common real-world re-ID system issues, e.g., similar-appearance identities, occlusions and misaligned frame. By capturing temporally static/dynamic and spatially coarse/fine information at different layers of a 3D-CNN, STRF produces robust discriminative representation to tackle these challenges as demonstrated here through attention maps of penultimate layer of feature extractor.

tracklets. With numerous applications in security, surveillance, and forensics [2], this problem has seen a dramatic increase in interest and various methodologies in the vision community [7, 13, 28, 29, 32, 36, 50, 54].

While there has been admirable progress in image-based re-ID as evidenced by recent quantitative results [7], there are many challenges that still preclude the ubiquitous use of re-ID algorithms in real-world systems. One such issue is appearance similarity, where multiple people wear similar looking clothes (e.g., large conferences or public events with a strict dress code). Other challenging issues include occlusions and frame misalignment that are a direct consequence of large crowd flow densities (e.g., in airports just after flight arrival) and inter-camera viewpoint disparities.

Having access to additional data, e.g., an extra temporal dimension like videos instead of 2D images, can help alleviate some of these issues by leveraging spatio-temporal data.

Video-based re-ID has seen much recent work [4, 5, 14, 19, 27, 45, 46, 49] in part due to the availability of relevant large-scale video datasets [41, 52]. However, learning a spatio-temporal representation that can alleviate the issues noted above still remains a challenge. While advances in general 3D convolutional networks (3D-CNNs) provide reasonable baseline spatio-temporal features, existing re-ID techniques typically rely on specialized architectures [20, 45, 46, 49] that are inflexible to be used with these baseline models. Other lines of work are focused entirely on learning either temporal or spatial representations separately [4, 5, 19], overlooking the complementarity that both streams of information provide in challenging scenarios, e.g., distinguishing people wearing similar clothes.

To address the aforementioned issues, we present a flexible new computational unit called Spatio-Temporal Representation Factorization (STRF) module. Given a feature volume from a certain 3D convolutional layer in a baseline 3D-CNN model, STRF extracts complementary information along both spatial ($h \times w$) and temporal (time, t) dimensions. By design, the proposed STRF module can be inserted in an existing 3D-CNN model after any convolutional layer, introducing only ∼0.15 million learnable parameters per unit (for instance, this results in only a ∼1.73% overall parameter increase with I3D [3]), resulting in a flexible and parameter-wise economic framework that is end-to-end trainable. STRF comprises two modules, called *temporal* feature factorization module (FFM) and *spatial* feature factorization module, to process feature tensors. The design principles of these modules are motivated by certain observations in video tracklets, which we discuss next.

The intuition behind STRF is demonstrated in Figure 1. We begin with the factorization module in the temporal dimension. First, the overall or "global" appearance of the person (e.g., color of clothes, skin, hair, etc) in a tracklet does not change (static) substantially over time. While one can argue these can change with illumination variations, we assume these variations are limited in a given camera view over a short period of time. Next, the walking patterns of a person may change over time, e.g., walking on a level surface *vs*. climbing stairs (dynamic). Consequently, there are two possible information factorization strategies when processing feature maps: low-frequency (static) sampling and high-frequency (dynamic) sampling. Low-frequency sampling of feature maps results in capturing the "slowly-moving" or approximately constant features, i.e., the appearance information. On the other hand, high-frequency sampling of feature maps results in capturing information that is more dynamically varying, i.e., walking patterns [31]. The temporal factorization module results in capturing static and dynamic features across time, which is especially helpful in identifying different individuals with similar appearance (see last row video tracklet in Figure 1).

The spatial factorization module, on the other hand, does the same low-frequency (which we call "coarse") and high-frequency (which we call "fine") sampling and processing as above, but along the spatial $h \times w$ dimensions. This is motivated by commonly occurring real-world issues such as occlusions and frame misalignment. Under these scenarios, the spatial FFM's high-frequency sampling and processing unit is able to capture more "details" of the person of interest as opposed to the other entities that are the causes of occlusion, or other background information in the case of misalignment. To understand this better, observe the attention maps for top row video tracklet in Figure 1. The baseline model, without our proposed module, highlights mostly the bicycle regions in the feature maps, whereas by adding our module, the model is able to capture the person regions in the frames more comprehensively. Similarly, to cover cases where there are no occlusions or misalignment, the spatial FFM's low-frequency sampling and processing unit become responsible for capturing more slowly-varying or spatially global appearance information. This results in the spatial factorization module to capture two separate streams of spatial information for robust representations.

To summarize, when multiple people in the gallery "look alike" (e.g., same clothes), features from our temporal factorization branch help disambiguate (i.e., people may look alike but walk differently). On the other hand, with occlusion/clutter, our idea is to rely on "local" features, which can be learned using our spatial branch. Our **main contributions** are as follows.

- We present a novel framework in video-based re-ID to learn discriminative 3D features by factorizing both temporal and spatial dimension of features into low-frequency (static/coarse) and high-frequency (dynamic/fine) components to tackle misalignment, occlusion, and similar appearance problems.

- To realize these factorization, we propose a flexible trainable unit with negligible computational overhead, called Spatio-Temporal Representation Factorization (STRF) module, that can be used in conjunction with any baseline 3D-CNN based re-ID architecture (see Figure 2).

- We conduct extensive experiments on multiple datasets to demonstrate how the proposed STRF module improves the performance of baseline architectures and also achieves state-of-the-art performance obtained by standard re-ID evaluation protocols (see Table 2 and 3).

2. Related Work

In this section, we review some recent methods pertaining to video-based person re-ID, and later discuss 3D-CNNs as feature extractors for video re-ID tasks.

Video-based re-ID. Following the success in image-based re-ID [7, 23, 28, 29, 32, 36, 50, 54], there has been much recent progress in video-based re-ID as well [5, 10, 14, 19,

27, 30, 44, 55]. For instance, [45] proposed multi-granular hypergraph learning framework which leveraged hierarchically divided feature maps at *last* layer of feature encoder with different levels of granularities to capture spatial and temporal cues, treating both the spatial and temporal dimensions the same. Additionally, there have also been a class of methods [19, 24] that perform feature modulation by *expanding* the feature extractor with additional learning modules instead of processing just the last layer's output as in [45]. Different from all the above works, we focus on learning factorized (dynamic/static and coarse/fine) information in **both** spatial and temporal dimensions (see Figure 2). This leads to a flexible feature processing module that can be used anywhere in any 3D-CNN based re-ID architecture, leading to improved performance of various baseline 3D-CNN models (see Table 2). We provide a characteristic comparison of recent works in Table 1.

3D-CNN based Feature Extractor. 3D-CNNs [21] naturally process input videos to output spatio-temporal features, whereas 2D-CNNs need additional modules such as recurrent networks to extract temporal information. Given this advantage, 3D-CNNs are more suitable for video-related applications [6, 9, 21, 42, 43], including video-based re-ID tasks [14, 25, 30]. For example, Gheissari *et al.* [11] introduced a two-stream model with the first branch comprised of 3D-CNNs and the other comprised of 2D-CNNs to extract temporal and spatial cues. In [14], appearance-preserving 3D convolution (AP3D) was proposed to leverage the idea of image registration [56] to perform feature-level image alignment. While these methods demonstrated good results, they either required both 3D and 2D CNNs [25], or additional operations, e.g., non-local convolutions, to achieve best performances [14], leading to parameter-wise bulky models. Furthermore, these methods do no explicitly exploit spatial cues of video tracklets. On the other hand, our proposed STRF method modifies the backbone feature encoder by means of a modular computational unit, does not require specialized modules such as recurrent networks or non-local operations, leading to only a minimal increase in learnable parameters while also demonstrating state-of-the-art performance on benchmark datasets.

Table 1: **Characteristic comparison with state-of-the-art works.** We compare our STRF with few current state-of-the-art works. Unlike these methods, STRF uses factorized information from *both* spatial (*S*) and temporal (*T*) dimensions, *does not require* non-local operations, and is *adaptable* to multiple baselines.

| METHODS | FACTORIZATION | | WITHOUT NON-LOCAL? | GENERIC? |
	T?	*S?*		
AP3D [14]	✗	✓	✗	✓
MGH [45]	✓	✓	✗	✗
AFA [5]	✓	✗	✓	✓
STRF (Ours)	✓	✓	✓	✓

3. Spatio-Temporal Factorization

As noted in Section 1, existing re-ID methods for learning video representations do not focus on the complementarity that is provided by the spatial and temporal dimensions. Specifically, we conjecture that the temporal dimension contains both static (e.g., appearance across time) as well as dynamic (e.g., walking patterns) content, whereas the spatial dimension comprises both fine (e.g., focus on details such as a person's legs that may be missed under occlusions) as well as coarse (e.g., overall global appearance) details. Consequently, we argue that all these features should be learned jointly in order to deal with unavoidable challenges such as appearance similarity, occlusions, and frame misalignment.

To address these issues, we introduce Spatio-Temporal Representation Factorization (STRF), a generic parameter-wise lightweight computational unit that can be inserted between convolutional layers in any 3D-CNN architecture for re-ID (note that by the term *factorization*, we refer to the joint sampling and processing operations for discussion below). This modularity makes STRF particularly appealing for practical applications that may require customized architectures based on data distribution. Along with the performance improvements in baseline architectures (see Table 2), STRF also demonstrates superior utility of the proposed module over existing specialized architectures for learning spatio-temporal re-ID representations (see Table 3) [14, 20, 25, 45, 46].

Notations. Let $\mathcal{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_t] \in \mathbb{R}^{t \times h \times w}$ denote an input video tracklet comprising t frames each of height h and width w. Let $\mathcal{F}_{\boldsymbol{\theta}}(\cdot)$ denote the feature encoder of any baseline 3D-CNN (e.g., I3D ResNet-50 [3]). Let $\boldsymbol{f}_\ell \in \mathbb{R}^{c_\ell \times t_\ell \times h_\ell \times w_\ell}$ be the feature tensor at the ℓth layer of $\mathcal{F}_{\boldsymbol{\theta}}(\cdot)$, where c_ℓ, t_ℓ, h_ℓ, and w_ℓ indicate number of channels, number of frames, height, and width, respectively. Let the input and output feature volumes of our STRF module at the ℓth layer be $\boldsymbol{f}_\ell^{(\boldsymbol{i})}$ and $\boldsymbol{f}_\ell^{(\boldsymbol{o})}$, respectively. Finally, let the static/coarse and dynamic/fine components be denoted with ς and τ, respectively and subscript $\boldsymbol{t}$ and $\boldsymbol{s}$ denote the temporal and spatial dimension, respectively. We use $d \in \{\boldsymbol{t}, \boldsymbol{s}\}$ and $k \in \{\tau, \varsigma\}$ for compact notations.

3.1. Feature Factorization Module (FFM)

Given $\boldsymbol{f}_\ell^{(\boldsymbol{i})}$, we propose to factorize this feature volume into four parts: static and dynamic content from temporal t_ℓ dimension, and coarse and fine detail from spatial $h_\ell \times w_\ell$ dimension. The intuition here is that the static content in the temporal dimension will capture "what does not change over time", e.g., appearance such as color of clothes, and the dynamic content will capture "what may change over time", e.g., walking patterns [31]. Similarly, coarse details in the spatial dimension will capture overall global information in the current feature map (e.g., "where is the person?") whereas fine detail helps address situations where the per-
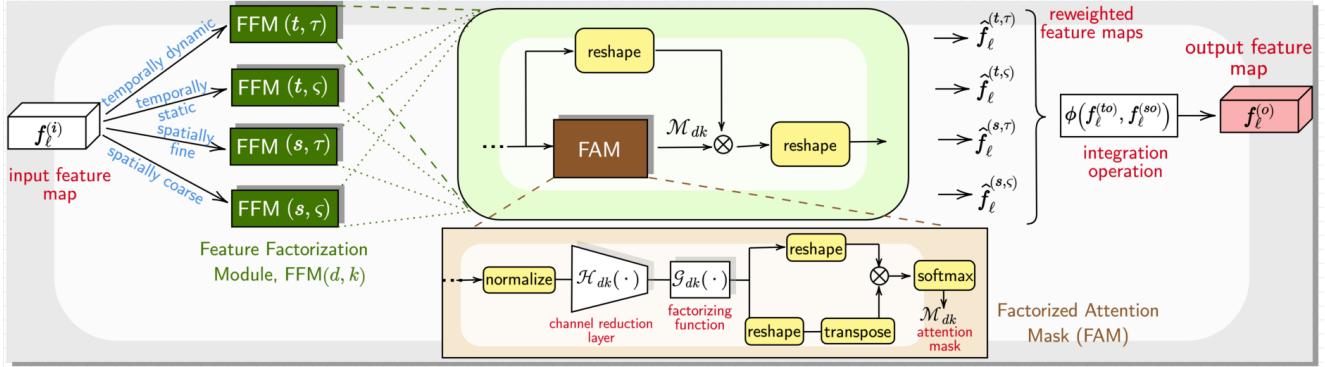
Figure 2: **Overview of STRF module**. Our proposed module contains four factorization units being applied on input feature volume at ℓth layer to extract static/coarse and dynamic/fine information and generate richer feature representation. Each unit is composed of a Feature Factorization Module aided by the proposed Factorized Attention Mask block.

son of interest may be occluded by other entities, by capturing local context at different locations of the feature map.

Our motivations above are particularly relevant given existing 3D-CNN architectures for re-ID do not have explicit mechanisms to focus on features corresponding to the person of interest in cases such as occlusions, image misalignment, or people with similar clothing appearing together in same tracklet. Furthermore, such a factorization enables a 3D-CNN to weight features that are important for downstream matching and re-ID, e.g., the dynamic content along the temporal dimension is more important in cases where people wear similar clothes and can be distinguished only by their walking patterns.

To realize this proposed factorization and feature re-weighting, STRF proposes to use four FFM modules, in which each FFM learns a different type of attention mask from **Factorized Attention Mask** (FAM) block (we discuss detailed architecture of FAM in next section) for either static/dynamic or coarse/fine content along the temporal and spatial dimensions respectively, and output refined feature volumes. Specifically, given $\boldsymbol{f}_\ell^{(i)}$, we first reshape it into the feature volume $\widehat{\boldsymbol{f}}_\ell^{(i)}$ with size $c_\ell t_\ell \times h_\ell w_\ell$ and then, use the FAM block to generate a factorized attention mask $\mathcal{M}_{dk}$. This mask is then used to compute a new feature volume as:

$$\widehat{\boldsymbol{f}}_\ell^{(dk)} = \widehat{\boldsymbol{f}}_\ell^{(i)} \mathcal{M}_{dk} \quad d \in \{\boldsymbol{t}, \boldsymbol{s}\}, k \in \{\tau, \varsigma\} \qquad (1)$$

STRF then integrates the four attention-weighted feature volumes $\{\widehat{\boldsymbol{f}}_\ell^{(t\tau)}, \widehat{\boldsymbol{f}}_\ell^{(t\varsigma)}, \widehat{\boldsymbol{f}}_\ell^{(s\tau)}, \widehat{\boldsymbol{f}}_\ell^{(s\varsigma)}\}$ to output a new feature volume which is then passed on to the subsequent convolutional layer. The output of this subsequent layer is then processed by the next instantiation of the STRF. This way, STRF provides a flexible computational unit that can be easily integrated with existing 3D-CNN architectures. Our proposed methodology is illustrated in Figure 2 where one can note that the four individual factorization modules, FFM$(\boldsymbol{t}, \tau)$, FFM$(\boldsymbol{t}, \varsigma)$, FFM$(\boldsymbol{s}, \tau)$, and FFM$(\boldsymbol{s}, \varsigma)$, combine to produce an enhanced feature representation $\boldsymbol{f}_\ell^{(o)}$ using their respective FAM blocks. We next discuss the fac-

torization attention masks and each of these proposed individual FFM modules in more detail.

3.2. Factorized Attention Masks (FAM) Block

To realize the four-way factorization for the feature volume $\boldsymbol{f}_\ell^{(i)}$ discussed above, we define four functions below:

$$\mathcal{T}_d^k\big(\boldsymbol{f}_\ell^{(i)}\big) = \mathcal{G}_{dk}\bigg(\mathcal{H}_{dk}\big(\boldsymbol{f}_\ell^{(i)}\big)\bigg), \qquad (2)$$

$$\text{with,} \quad d \in \{\boldsymbol{t}, \boldsymbol{s}\}, k \in \{\tau, \varsigma\}$$

where $\mathcal{G}_{dk}(\cdot)$ are the **factorizing functions**. Different for each FFM block, $\mathcal{G}_{dk}(\cdot)$ is designed using pooling functions to extract specific information after the input feature volume is passed through a channel reduction layer $\mathcal{H}_{dk}(\cdot) : c_\ell \to c_\ell/n$, where $\mathcal{H}_{dk}(\cdot)$ is a convolutional layer with c_ℓ/n kernels of size 1. Following [14, 22], we set $n = 16$. With the output composite function $\mathcal{T}_d^k\big(\boldsymbol{f}_\ell^{(i)}\big)$ of size $c_\ell/nt_\ell \times h_\ell w_\ell$ from (2), we summarize input features by computing their variance matrix $\mathcal{C}_{dk}$ to obtain a representation of each point of $\mathcal{T}_d^k\big(\boldsymbol{f}_\ell^{(i)}\big)$ as:

$$\mathcal{C}_{dk} = \kappa \mathcal{T}_d^k\big(\boldsymbol{f}_\ell^{(i)}\big)^\top \mathcal{T}_d^k\big(\boldsymbol{f}_\ell^{(i)}\big) \qquad (3)$$

where $\top$ represents transpose operation. We set the temperature hyper-parameter κ as 4 following [14, 19]. Then, the factorizing mask is computed using the unnormalized sample covariance matrix as $\mathcal{M}_{dk}(q) = \sigma(\mathcal{C}_{dk})$, where $\sigma(\cdot)$ is the softmax function. This factorized mask is employed in (1) to obtain the specific factorized representation of $\boldsymbol{f}_\ell^{(i)}$. Next, each factorization module is discussed in more detail.

Temporal Factorization Module, FFM$(\boldsymbol{t}, \tau, \varsigma)$. While methods for learning static and dynamic information have been presented in prior work [1, 9, 16, 38], we take a modular approach to this problem, proposing computational units that can be applied at multiple layers of the base feature encoder. Instead of skipping frames as in [9], we define the
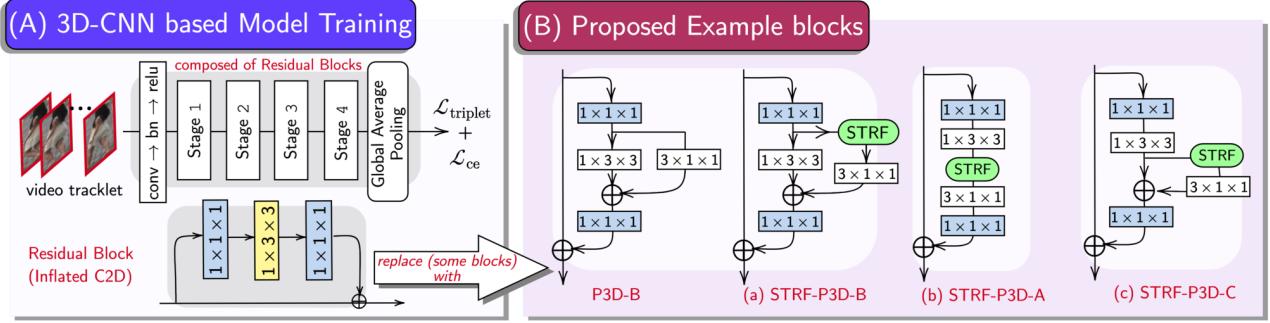
Figure 3: **Illustration of our 3D-CNN model training and examples of proposed blocks**. In our framework, we employ a 3D-CNN based model to learn discriminative features for input video tracklets (see Figure (A)). This model is built with inflated 2D-residual blocks with Stage 2 and Stage 3 replaced by our proposed STRF aided residual blocks (see Figure (B)).

following temporal factorizing functions:

$$\mathcal{G}_{t\tau} = \texttt{pool}\big(r_{t\tau}, 1, 1\big),\ \mathcal{G}_{t\varsigma} = \texttt{pool}\big(r_{t\varsigma}, 1, 1\big) \quad (4)$$

where $r_{t\varsigma} > r_{t\tau}$. These degenerate `pool` functions can be implemented using the *max pooling* (denoted as m) and *average pooling* (denoted as a) operations with their corresponding static temporal resolutions $r_{t\varsigma}$ and dynamic temporal resolutions $r_{t\tau}$. We also use suitable padding on $\mathcal{H}_{t\varsigma}(\boldsymbol{f}_\ell^{(i)})$ and $\mathcal{H}_{t\tau}(\boldsymbol{f}_\ell^{(i)})$ to maintain the same size between the input and output feature volumes. The intuition behind setting $r_{t\varsigma} > r_{t\tau}$ is to factorize features in time dimension to capture the information that does not vary much, whereas $r_{t\tau}$ helps in summarizing information that shows more variations. Capturing such static information with $\mathcal{G}_{t\varsigma}$ will aid in learning the global appearance features of the person that does not change much along the time dimension. On the other hand, $\mathcal{G}_{t\tau}$ captures dynamic information in the input feature volume, e.g., walking patterns of the person. Finally, the output of FFM$\big(\boldsymbol{t}, \tau, \varsigma\big)$ is defined as:

$$\boldsymbol{f}_\ell^{(to)} = \widehat{\boldsymbol{f}}_\ell^{(t\tau)} + \widehat{\boldsymbol{f}}_\ell^{(t\varsigma)} \quad (5)$$

where $\widehat{\boldsymbol{f}}_\ell^{(t\tau)}$ and $\widehat{\boldsymbol{f}}_\ell^{(t\varsigma)}$ are computed using (1).

Spatial Factorization Modules, FFM$\big(\boldsymbol{s}, \tau, \varsigma\big)$. Similar to the temporal dimension above, we factorize the feature volume along the spatial dimension as well, extracting coarse-level and fine-level information. The intuition here is that coarse-level information in the spatial dimension comprise global features of the person in the input frames that do not have much occlusion. For frames where the person is occluded or there is spatial misalignment, fine-level features capture the "person-part" of the frame. To realize this, we define the following spatial factorizing functions:

$$\mathcal{G}_{s\tau} = \texttt{pool}\big(1, r_{s\tau}, r_{s\tau}\big),\ \mathcal{G}_{s\varsigma} = \texttt{pool}\big(1, r_{s\varsigma}, r_{s\varsigma}\big) \quad (6)$$

where $r_{s\varsigma} > r_{s\tau}$ are the spatially coarse and fine resolution, respectively. As in FFM$\big(\boldsymbol{t}, \tau, \varsigma\big)$, we use appropriate padding on $\mathcal{H}_{s\varsigma}(\boldsymbol{f}_\ell^{(i)})$ and $\mathcal{H}_{s\tau}(\boldsymbol{f}_\ell^{(i)})$ to maintain the same

size between the input and output feature volumes. Finally, the output of FFM$\big(\boldsymbol{s}, \tau, \varsigma\big)$ is defined as:

$$\boldsymbol{f}_\ell^{(so)} = \widehat{\boldsymbol{f}}_\ell^{(s\tau)} + \widehat{\boldsymbol{f}}_\ell^{(s\varsigma)} \quad (7)$$

where $\widehat{\boldsymbol{f}}_\ell^{(s\tau)}$ and $\widehat{\boldsymbol{f}}_\ell^{(s\varsigma)}$ are computed using (1). Note that when the resolutions are set as 1 in (6) and (4), the factorizing functions behave as identity mapping. In our experiments, we set $r_{s\tau} = r_{t\tau}$ and $r_{s\varsigma} = r_{t\varsigma}$ for simplicity.

Integration and overall STRF output. After computing $\boldsymbol{f}_\ell^{(to)}$ and $\boldsymbol{f}_\ell^{(so)}$ as discussed above, we provide two schemes to integrate them and generate the final feature volume output of our proposed STRF computational unit:

$$\boldsymbol{f}_\ell^{(o)} = \phi\big(\boldsymbol{f}_\ell^{(to)}, \boldsymbol{f}_\ell^{(so)}\big) \quad \text{where,}\ \phi(\,\cdot\,) \in \{\rightarrow, \|\} \quad (8)$$

Here, $\rightarrow$ denotes using the temporal and spatial factorization modules in cascade, and $\|$ denotes using them in parallel. When in cascade, the input $\boldsymbol{f}_\ell^{(i)}$ is fed to both modules in sequence, *i.e.* FFM$\big(\boldsymbol{s}, \tau, \varsigma\big)$ followed by FFM$\big(\boldsymbol{t}, \tau, \varsigma\big)$, or *vice-versa*. When in parallel, the outputs of FFM$\big(\boldsymbol{s}, \tau, \varsigma\big)$ and FFM$\big(\boldsymbol{t}, \tau, \varsigma\big)$ are simply added. In our experiments, we noticed only minor performance differences across these operations (see Figure 4(c)).

Learning Objective. Any STRF-aided network can be trained in an end-to-end manner with following objective:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{triplet}} \quad (9)$$

where $\mathcal{L}_{\text{ce}}$ is the standard cross-entropy classification, $\mathcal{L}_{\text{triplet}}$ is the cosine distance based triplet loss with batch-hard mining [18], and $\mathcal{L}$ is the overall loss function. Note that our method demonstrates state-of-the-art results (see Table 3) without any re-ID tricks [33], e.g. label smoothing [37], in our learning objective.

How do we employ STRF? The problem of person re-ID benefited tremendously with introduction of residual blocks [12, 17]. With the backbone feature extractor as inflated
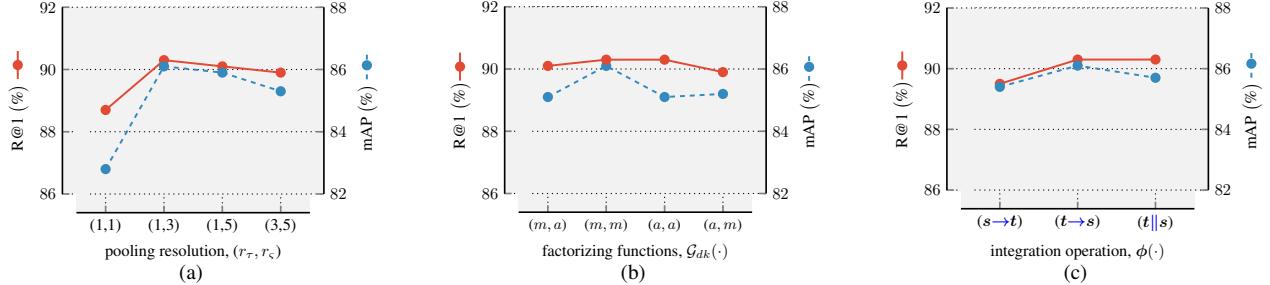
Figure 4: **Analysis of different components of STRF**. (a) Each $(r_\varsigma, r_\varsigma)$ refers to spatially coarse resolutions : $(1, r_\varsigma, r_\varsigma)$, spatially fine resolutions : $(1, r_\tau, r_\tau)$, temporally static resolutions : $(r_\varsigma, 1, 1)$, temporally dynamic resolutions : $(r_\tau, 1, 1)$. Best results are obtained with $(r_\varsigma, r_\varsigma) = (1, 3)$. (b) Performance of different combinations of factorizing functions $\mathcal{G}_{dk}(\cdot)$: Best results are obtained when all $\mathcal{G}_{dk}(\cdot)$ are set as maxpooling function. (c) Performance of different integration operations $\phi(\cdot)$: Best results are obtained when the spatial module is followed by the temporal module.

C2D (time dimension of kernel set to 1) residual network, we propose to enhance its feature representation learning paradigm by simply replacing residual blocks at different stages with different STRF-aided I3D or STRF-aided Pseudo-3D (P3D) [34] residual blocks (see Figure 2(B)). To convert P3D residual blocks to their STRF-P3D forms, we add the STRF module with the convolutional layer of kernel size $3 \times 1 \times 1$ demonstrating the generic ability of the proposed unit. We have empirically analyzed and discussed this choice of location in the supplementary material. Moreover, a single STRF module introduces only minimal extra-parameters which makes it parameter-wise lightweight but performance-wise beneficial (see Table 2).

3.3. Discussion

FAM vs Channel Attention (CA). We note that there are substantial differences between FAM and the popular CA strategy [8, 15, 48]. Unlike CA that has one *global* feature pooling layer, i.e., no separate spatial and temporal operations, FAM has **four** pooling functions $\mathcal{G}_{dk}(\cdot)$, defined in (4) and (6). This captures both spatial and temporal feature dependencies without *any* new learning parameters. In fact, with r_ς and r_τ set to same size of the input feature maps, CA can be considered to be a special case of FAM.

FFM vs Non-Local (NL). Unlike the popular NL module [40] where there is *no* factorization, FFM factorizes $f^{(i)}$ into its constituent spatio/temporal factors. The appropriate weighting of $f^{(i)}$ with these factors to obtain $f^{(o)}$ is automatically learned with FAM, making the proposed design different from NL and more suitable for re-ID. For additional empirical substantiation, using the P3DC architecture on the MARS dataset [52], the NL module gives 84.8% mAP and 89.9% R@1, whereas STRF gives 86.1% mAP and 90.3% R@1. Further, STRF only adds an additional $\sim$0.5 million parameters (w.r.t. the baseline) as opposed to NL's $\sim$5 million additional parameters, demonstrating better compute efficiency.

Please see supplementary material for additional insights and discussions on our proposed STRF module.

4. Experimentation

Datasets, implementation details, and evaluation metrics. We conduct extensive experiments on standard publicly available video-based person re-ID datasets, including MARS [52], DukeMTMC-VideoReID [41], and iLIDS-VID [39]. For evaluation, we use the value of the cumulative matching characteristic curve at rank-1 (R@1), and mean average precision (mAP) [53]. See supplementary material for full implementation details.

4.1. Improvement over Baselines

Quantitative analysis. We build a model with inflated 2D convolutions in ResNet50 (temporal kernel size set to 1) architecture. We then replace stage 2 and stage 3 (See Table 5) with four residual blocks **I3D** (temporal kernel size set to 3) and three pseudo-3D residual blocks **P3D-A**, **P3D-B** and **P3D-C** to create four baselines. For comparative evaluation, we replace these I3D and P3D residual blocks with **STRF-I3D**, **STRF-P3DA**, **STRF-P3DB** and **STRF-P3DC** residual blocks respectively and summarize the results in Table 2. One can clearly note that the STRF-aided models give improved performance (at least 2.5% mAP increment for P3D baselines and about 0.5% mAP increment for I3D baseline on MARS), with the best performance achieved with STRF-P3DC. Similar trends can be observed on the DukeMTMC-VideoReID as well. Furthermore, when compared to the number of baseline model parameters (denoted in Table 2 as **P**(M) on MARS in the millions of parameters), the number of new parameters introduced by our proposed module is only 0.05 million more compared with I3D or P3D models, suggesting it does not add any substantial computational overhead. This also demonstrates that STRF can improve performance of *diverse* architectures. For all subsequent experiments, we report results with STRF-P3DC following its best performance from Table 2.

Qualitative Analysis. To qualitatively demonstrate STRF's impact, we visualize feature maps of challenging videos (e.g., occlusions, misalignment) using off-the-shelf techniques [14, 47] in Figure 5. Note that STRF helps

Table 2: **Baseline improvement.** STRF consistently improves the performance of baseline models. **P**(M) is model size in millions.

MODEL	P(M)	DATASETS			
		MARS [52]		DukeMTMC [41]	
		mAP (%)	R@1 (%)	mAP (%)	R@1 (%)
I3D	28.92	82.70	88.50	95.20	95.40
+ STRF	28.97	**83.10**	**88.70**	**95.20**	**95.90**
P3DA	25.48	83.20	88.90	95.00	95.00
+ STRF	25.53	**85.40**	**89.80**	**95.60**	**96.00**
P3DB	25.48	83.00	88.80	95.40	95.30
+ STRF	25.53	**85.60**	**90.30**	**96.40**	**97.40**
P3DC	25.48	83.10	88.50	95.30	95.30
+ STRF	25.53	**86.10**	**90.30**	**96.20**	**97.20**

focus more clearly on the person of interest (e.g., under "occlusion", unlike the baseline, STRF is able to more clearly distinguish between person's foreground and occlusion regions). Please see supplementary material for more qualitative and attention map results.

4.2. Ablation Study

Utility of FAM block. Our temporal and spatial factorization modules are realized with the proposed factorized attention masks $\mathcal{M}_{dk}$. These self-attention masks are utilized to re-weight the input feature volume $\boldsymbol{f}_\ell^{(i)}$ in order to produce a richer representation of the video tracklet. Specific information captured via $\mathcal{M}_{dk}$ (due to different $\mathcal{G}_{dk}$ for both low-frequency (static/coarse) and high-frequency (dynamic/fine) information) enhance input feature volume to represent robust features by re-weighting them as in (1). Consequently, FAM is an important component of our proposed STRF module. To validate this, we present an analysis of STRF with and without the FAM in Figure 6(a) on MARS [52]. It can be observed that without FAM, the proposed module weakens the feature representations (non-weighted multiplication ($\otimes$) of $\boldsymbol{f}^{(i)}$ with itself) resulting in a comparatively lower performance. More concretely, without FAM, we do not have "coarse/fine" and "static/dynamic" factors, and hence FFM does not receive appropriate factors to re-weight $\boldsymbol{f}^{(i)}$. Note that using FAM alone (without FFM) is not possible by design.

Analysis of different components. The proposed static and dynamic factorizing functions differ essentially in their temporal and spatial resolutions, and in Figure 4(a), we analyze various combinations of these resolution parameters while keeping the static resolution (or coarser) r_ς larger than the dynamic (or finer) resolution r_τ. Note that we keep the low- and high-frequency (static and coarse) resolutions of both these modules the same for simplicity and reduced parameter search space. The maximum resolution is dependent on output size of the last conv layer STRF is applied to. In our case, this last layer output is $\boldsymbol{f}^{(i)} \in \mathbb{R}^{2048 \times 8 \times 14 \times 7}$, giving only possible choices of 1, 3, 5, and 7. A (1, 7, 7) filter will give $\boldsymbol{f}^{(o)} \in \mathbb{R}^{2048 \times 8 \times 7 \times 1}$, i.e., 7×1 spatial dimension, unsuitable for computing $\mathcal{M}_{dk}$. As coarse resolution should be larger than fine resolution, only 4 plausible
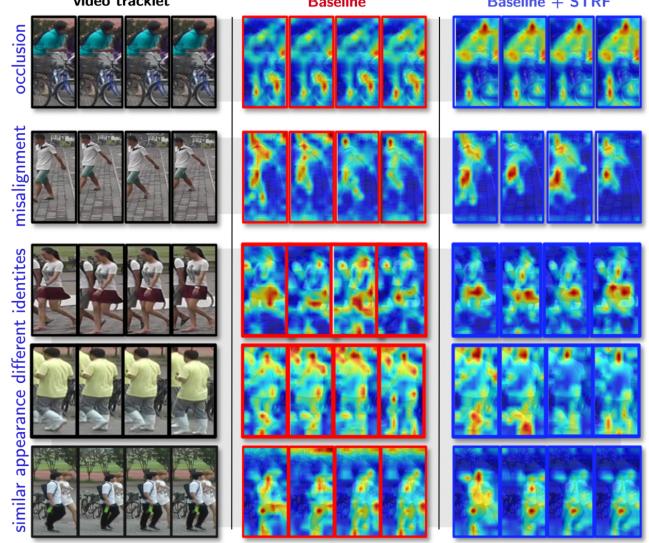

Figure 5: **Attention map visualizations.** STRF helps baseline models extract more discriminative features.

pairs (including $(r_\tau, r_\varsigma) = (1, 1)$ for reference) results are presented. One can note from the graph that STRF performs the best with the resolution pair $(r_\tau, r_\varsigma) = (1, 3)$. The graph also shows that STRF is not very sensitive to the different resolution pairs, with a difference of 0.4% mAP when $(r_\tau, r_\varsigma) = (3, 5)$, and difference of 0.2% mAP when $(r_\tau, r_\varsigma) = (1, 5)$. Next, we analyze various combinations of factorizing functions defined as part of STRF modules in Figure 4(b). Our framework performs the best with both temporal and spatial $\mathcal{G}_{dk}$ set to the *max pool* (*m*) operation. This is likely because factorization based on *max pooling* helps focus on information that represents the discriminative portion of the feature volume. Finally, we analyze the different integration functions described in (8), where we note that the best performance is obtained when we first factorize $\boldsymbol{f}_\ell^{(i)}$ by the temporal factorization module FFM($\boldsymbol{t}, \tau, \varsigma$) and then feed this output to spatial factorization module FFM($\boldsymbol{s}, \tau, \varsigma$), i.e., when $\phi(\cdot) = \rightarrow$. Further, when $\phi(\cdot) = \parallel$, a comparable performance is observed with a difference of about 0.4% in mAP.

Which stage to add? Table 5 presents results of adding STRF at various stages of a baseline model. Using STRF module in Stage 2 and Stage 3 gives the best performance, but reduces (in mAP) when added to Stage 1. This is likely because with Stage 1, low-level features do not contain enough descriptive semantic information for detailed factorization. Additionally, Stage 4 (last two rows in Table 5) exhibits differing behavior, likely due to the feature pooling operation performed at this layer (for subsequent classification), which provides spatio-temporally-entangled gradients which may not be useful for our STRF module. Please see supplementary material for more results.

Influence of each factorization module. To study the efficacy of each module, we perform an ablation analysis (see

Table 3: **Comparison with the state-of-the-art.** STRF gives state-of-the-art performance on all datasets (best results in red, second best in blue, and third best results in green.)

| METHODS | VENUE | DATASETS | | | | |
| | | MARS [52] | | DukeMTMC [41] | | iLiDS-VID [39] |
		mAP (%)	R@1 (%)	mAP (%)	R@1 (%)	R@1 (%)
ADFD [51]	CVPR 2019	78.20	87.00	–	–	86.30
VRSTC [20]	CVPR 2019	82.30	88.50	93.50	95.00	86.30
COSAM [35]	ICCV 2019	79.90	84.90	94.10	95.40	79.60
GLTR [24]	ICCV 2019	78.50	87.00	93.74	96.29	86.00
MGH [45]	CVPR 2020	85.80	90.00	–	–	85.60
STGCN [46]	CVPR 2020	83.70	89.95	95.70	97.29	–
MG-RAFA [49]	CVPR 2020	85.90	88.80	–	–	88.60
TACAN [26]	WACV 2020	84.00	89.10	95.40	96.20	88.90
M3D [25]	TPAMI 2020	79.46	88.63	93.67	95.49	86.67
AFA [5]	ECCV 2020	82.90	90.20	95.40	97.20	88.50
AP3D [14]	ECCV 2020	85.60	90.70	96.10	97.20	88.70
TCLNet [19]	ECCV 2020	85.10	89.80	96.20	96.90	86.60
STRF	**Ours**	**86.10**	**90.30**	**96.40**	**97.40**	**89.30**

Table 4). Each individual module FFM(t, τ), FFM(t, ς), FFM(s, τ), and FFM(s, ς) improves the baseline with at least 2% in mAP and 1.2% in R@1. Further, temporal and spatial factorization modules perform better when used together. The temporal/spatial similarity (in margins) in Table 4 suggests each module is equally effective in identifying unique features w.r.t. baseline. Finally, the best performance is obtained when all modules are put together, demonstrating their focus on complementary information.

4.3. Comparison with state-of-the-art approaches

Despite being parameter-wise lightweight and agnostic to baseline architectures, STRF gives competitive results

Table 4: **Contribution of each factorization module.** All four STRF modules FFM(t, τ), FFM(t, ς) FFM(s, τ), and FFM(s, ς) show improvement individually and collectively with the P3DC baseline on MARS [52].

| MODEL | MODULES | | | | MARS [52] | |
	(s, τ)	(s, ς)	(t, τ)	(t, ς)	mAP (%)	R@1 (%)
Baseline					83.10	88.50
Baseline + STRF	✓				85.20	89.70
		✓			85.10	89.90
			✓		85.20	89.90
				✓	85.10	90.00
	✓	✓			85.50	90.10
			✓	✓	85.30	89.70
	✓		✓		85.40	90.00
		✓		✓	85.70	90.10
	✓	✓	✓	✓	**86.10**	**90.30**

Table 5: **Per-stage influence of STRF.** All four STRF modules are effective at various stages, with best results at Stage 2 and 3 of STRF-P3DC on MARS [52].

MODEL	STAGE	mAP (%)	R@1 (%)
Baseline		83.10	88.50
Baseline + STRF	1	83.40	88.80
	1, 2	83.60	89.00
	2, 3	**86.10**	**90.30**
	1, 2, 3	84.70	89.30
	2, 3, 4	85.50	90.00
	1, 2, 3, 4	83.70	88.70

when compared to sophisticated 3D-CNN methods. As can be observed in Figure 6(b), STRF outperforms both AP3D and M3D with ∼6 million (w.r.t. AP3D [14]) and ∼75 million (w.r.t. M3D [25]) **fewer** parameters. Finally, STRF establishes a new state-of-the-art (w.r.t. mAP) on MARS, DukeMTMC, and iLIDS-VID as shown in Table 3.
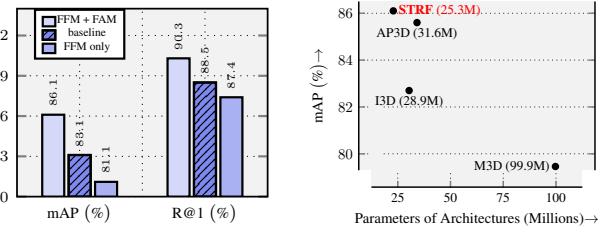

(a) FAM ablation analysis (b) Parameters *vs* mAP analysis

Figure 6: **Advantages of STRF**. (a) Without FAM block, FFM cannot factorize features leading to poor performance, signifying FAM's importance in STRF. (b) STRF is comparatively parameter-wise most light-weight and best performing 3D-CNN architecture.

5. Conclusion

We proposed a novel Spatio-Temporal Representation Factorization (STRF) computational unit that learns complementary spatio-temporal feature representations to deal with real-world re-ID challenges such as occlusions, imperfect detection, and appearance similarity. Our STRF module factorizes temporal dynamic/static, and spatial coarse/fine components from input 3D-CNN feature maps, helping baseline models discover more complementary and discriminative spatio-temporal representations for robust video re-ID. Extensive evaluations of our STRF module with various baseline architectures on benchmark video-based re-ID datasets show its efficacy and generality. As part of future work, we would like to extend it to general video understanding problems like semantic segmentation.

References

[1] Abhishek Aich, Akash Gupta, Rameswar Panda, Rakib Hyder, M Salman Asif, and Amit K Roy-Chowdhury. Non-Adversarial Video Synthesis with Learned Prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6090–6099, 2020. 4

[2] Octavia Camps, Mengran Gou, Tom Hebble, Srikrishna Karanam, Oliver Lehmann, Yang Li, Richard J Radke, Ziyan Wu, and Fei Xiong. From the Lab to the Real World: Re-Identification in an Airport Camera Network. *IEEE Transaction on Circuit Systems and Video Technology*, 27(3):540–553, 2016. 1

[3] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2, 3

[4] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 1169–1178, 2018. 2

[5] Guangyi Chen, Yongming Rao, Jiwen Lu, and Jie Zhou. Temporal Coherence or Temporal Motion: Which is More Critical for Video-based Person Re-identification? In *Proceedings of the European Conference of Computer Vision*, 2020. 2, 3, 8

[6] Wei Chen, Boqiang Liu, Suting Peng, Jiawei Sun, and Xu Qiao. S3D-UNet: Separable 3D U-Net for Brain Tumor Segmentation. In *International MICCAI Brainlesion Workshop*, pages 358–368. Springer, 2018. 3

[7] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, and Y. Yang. Salience-Guided Cascaded Suppression Network for Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2

[8] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-Order Attention Network for Single Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 6

[9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019. 3, 4

[10] Jiyang Gao and Ram Nevatia. Revisiting Temporal Modeling for Video-based Person Re-Id. *arXiv preprint arXiv:1805.02104*, 2018. 2

[11] N. Gheissari, T. B. Sebastian, and R. Hartley. Person Re-identification Using Spatio-temporal Appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1528–1535, 2006. 3

[12] Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, Richard J Radke, et al. A Systematic Evaluation and Benchmark for Person Re-Identification: Features, Metrics, and Datasets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):523–536, 2018. 5

[13] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating Appearance Models for Recognition, Re-acquisition, and Tracking. In *Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, 2007. 1

[14] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-Preserving 3D Convolution for Video-based Person Re-identification. In *Proceedings of the European Conference of Computer Vision*, 2020. 2, 3, 4, 6, 8

[15] Akash Gupta, Abhishek Aich, and Amit K Roy-Chowdhury. ALANET: Adaptive Latent Attention Network for Joint Video Deblurring and Interpolation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 256–264, 2020. 6

[16] Akash Gupta, Padmaja Jonnalagedda, Bir Bhanu, and Amit K. Roy-Chowdhury. Ada-VSR: Adaptive Video Super-Resolution with Meta-Learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2021. 4

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[18] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737*, 2017. 5

[19] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal Complementary Learning for Video Person Re-Identification. In *Proceedings of the European Conference of Computer Vision*, 2020. 2, 3, 4, 8

[20] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. VRSTC: Occlusion-Free Video Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7176–7185. IEEE, 2019. 2, 3, 8

[21] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2012. 3

[22] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style Normalization and Restitution for Generalizable Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3143–3152, 2020. 4

[23] Srikrishna Karanam, Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, Richard J Radke, et al. A Systematic Evaluation and Benchmark for Person Re-Identification: Features, Metrics, and Datasets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):523–536, 2018. 2

[24] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-Local Temporal Representations for Video Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3958–3967, 2019. 3, 8

[25] J. Li, S. Zhang, and T. Huang. Multi-scale temporal cues learning for video person re-identification. *IEEE Transactions on Image Processing*, 29:4461–4473, 2020. 3, 8

[26] Mengliu Li, Han Xu, Jinjun Wang, Wenpeng Li, and Yongli 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. 8

[27] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity Regularized Spatiotemporal Attention for Video-based Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Recognition, pages 369–378, 2018. 2, 3

[28] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious Attention Network for Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[29] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person Re-Identification by Local Maximal Occurrence Representation and Metric Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2197–2206, 2015. 1, 2

[30] Xingyu Liao, Lingxiao He, Zhouwang Yang, and Chi Zhang. Video-based Person Re-Identification via 3D Convolutional Networks and Non-Local Attention. In *Asian Conference on Computer Vision*, pages 620–634, 2018. 3

[31] Yasushi Makihara, Daisuke Adachi, Chi Xu, and Yasushi Yagi. Gait Recognition by Deformable Registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 561–571, 2018. 2, 3

[32] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual Attention Networks for Multimodal Reasoning and Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 299–307, 2017. 1, 2

[33] Priyank Pathak, Amir Erfan Eshratifar, and Michael Gormish. Video Person Re-ID: Fantastic Techniques and Where to Find Them. *arXiv preprint arXiv:1912.05295*, 2019. 5

[34] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5533–5541, 2017. 6

[35] Arulkumar Subramaniam, Athira Nambiar, and Anurag Mittal. Co-Segmentation Inspired Attention Networks for Video-based Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 562–572, 2019. 8

[36] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In *Proceedings of the European Conference of Computer Vision*, 2018. 1, 2

[37] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 5

[38] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing Motion and Content for Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1526–1535, 2018. 4

[39] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person Re-Identification by Video Ranking. In *Proceedings of the European Conference on Computer Vision*, pages 688–703. Springer, 2014. 6, 8

[40] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 6

[41] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the Unknown Gradually: One-Shot Video-based Person Re-Identification by Stepwise Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2018. 2, 6, 7, 8

[42] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking Spatio-Temporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification. In *Proceedings of the European Conference on Computer Vision*, pages 305–321, 2018. 3

[43] Huijuan Xu, Abir Das, and Kate Saenko. R-C3D: Region Convolutional 3D Network for Temporal Activity Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5783–5792, 2017. 3

[44] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly Attentive Spatial-Temporal Pooling Networks for Video-based Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4733–4742, 2017. 3

[45] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, and Ling 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. 2, 3, 8

[46] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi 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. 2, 3, 8

[47] Sergey Zagoruyko and Nikos Komodakis. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. *arXiv preprint arXiv:1612.03928*, 2016. 6

[48] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image Super-Resolution using Very Deep Residual Channel Attention Networks. In *Proceedings of the European Conference on Computer Vision*, pages 286–301, 2018. 6

[49] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo 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. 2, 8

[50] 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*, 2020. 1, 2

[51] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xiansheng Hua. Attribute-Driven Feature Disentangling and Temporal Aggregation for Video Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4913–4922, 2019. 8

[52] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. MARS: A Video Benchmark for Large-Scale Person Re-Identification. In *Proceedings of the European Conference on Computer Vision*, pages 868–884, 2016. 2, 6, 7, 8

[53] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable Person Re-Identification: A Benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1116–1124,

2015. 6

[54] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke. Re-Identification with Consistent Attentive Siamese Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[55] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 4747–4756, 2017. 3

[56] Barbara Zitova and Jan Flusser. Image Registration Methods: A Survey. *Image and Vision Computing*, 21(11):977–1000, 2003. 3