Fine-Grained Re-Identification

Priyank Pathak
New York University
ppriyank@nyu.edu

Abstract

Research into the task of re-identification (ReID) is picking up momentum in computer vision for its many use cases and zero-shot learning nature. This paper proposes a computationally efficient fine-grained ReID model, FGReID, which is among the first models to unify image and video ReID while keeping the number of training parameters minimal. FGReID takes advantage of video-based pre-training and spatial feature attention to improve performance on both video and image ReID tasks. FGReID achieves state-of-the-art (SOTA) on MARS, iLIDS-VID, and PRID-2011 video person ReID benchmarks. Eliminating temporal pooling yields an image ReID model that surpasses SOTA on CUHK01 and Market1501 image person ReID benchmarks. The FGReID achieves near SOTA performance on the vehicle ReID dataset VeRi as well, demonstrating its ability to generalize. Additionally we do an ablation study analyzing the key features influencing model performance on ReID tasks. Finally, we discuss the moral dilemmas related to ReID tasks, including the potential for misuse. Code for this work is publicly available at https://github.com/ppriyank/Fine-grained-Re-Identification.

1. Introduction

New security systems and smart traffic grids require robust algorithms that can re-identify similar objects (faces, people, vehicles, etc.) across different camera viewpoints. Re-identification (ReID) aims to match identical objects such as people (person ReID), vehicles (vehicle ReID) and faces (face ReID), that experience subtle variations across viewpoints and time. ReID is most useful for identifying unique objects for surveillance purposes. However, ReID can also be used to create embeddings or indices for generic search engines. While image ReID depends on a single frame for re-identification and is susceptible to complexities such as occlusion, low illumination, and inferior viewpoints, video ReID uses multiple frames across time to counteract any identification impediments in a single image.

The downside of video ReID over image ReID is the added complexity of handling additional temporal information. Existing person ReID works have utilized graph convolutional neural networks [37, 42, 43], self-attention [15, 26], temporal attention [6, 13, 16], and LSTMs [2]. Many of these works depend on the appearance of objects (spatial structure), overlooking the fine-grained subtleties that may contain important distinguishing attributes. For example, existing approaches often fail when people have similar clothing styles, as demonstrated in Figure 1. The aforementioned methods are computationally expensive and impose restrictions on the input image size and batch size, critical to performance improvement. Numerous image ReID models cannot be generalized to videos and vice versa, discarding surplus hours of available CCTV (Closed-circuit television (Surveillance camera)) footage. We propose a unified approach for video and image ReID that requires fewer training parameters and uses fine-grained details to re-identify individuals more accurately.

The task of learning fine-grained details can be divided into implicit and explicit methodologies. Implicit methods include parts-based matching [3, 4, 34], which are often contingent on the alignment of parts, and feature erasing [44] to indirectly learn minute details. Explicit methods [11, 46] commonly deploy spatial attention maps to isolate and highlight subtle details. While such methodologies perform very well for images, they are not scalable for long videos. In this work, we extend an explicit fine-grained image classification technique [56] to videos by employing an extra ResNet network for only handing minute details.
Video person ReID models must be capable of handling both temporal and spatial variation. Numerous works [6, 13, 16] use a specialized temporal attention model to merge structures across frames. Temporal attention models often assign smaller weights to frames with heavy occlusion, but that may cause loss of distinctive features. Fu et al. [5] argue that occluded frames still contribute characteristic details like those in Figure 2. Su et al. [33] demonstrated some of the non-occluded outlier frames might confuse a model in targeting the wrong individuals. Figure 3 shows the red highlighted frames focusing on the black-dressed individual compared to the correct individual identified by green highlighted frames. In such cases, temporal attention fails, favoring computationally cheap temporal pooling. We also include non-local operations [39] to introduce context that can disregard outlier frames.

Our contributions are as follows. 1- We adopt a fine-grained image classification model and propose a novel framework, FGReID, capable of generating contextually-aware fine-grained embeddings for images and videos. It’s ability to re-identify is not only limited to people but vehicles as well. FGReID is among a few frameworks that perform equally well on images and videos, making it optimal for a wide range of ReID problems. 2- The model employs a limited number of training parameters making it computationally inexpensive and lightweight while capturing fine-grained details in one pass. We show FGReID size and computation time comparison with publicly available video ReID models. 3- Extensive experiments show FGReID exceeding SOTA on two large-scale video person ReID datasets MARS and iLIDS-VID while matching SOTA on the PRID-2011 dataset. For image ReID, our model exceeds SOTA on Market1501, CUHK01, while it is on par with SOTA on the VeRi-776 dataset. 4- We also address the ethical concerns regarding ReID work.

2. Related Work

Image ReID: Recent image person ReID work either relies on deep learning models [26, 28, 41] or on techniques [12, 27, 30, 36, 52, 54] for boosting the performance of existing frameworks. Recently, Liu et al. [20] utilized generative adversarial networks (GAN) to produce style invariant images from various camera viewpoints. GANs often blur pictures in recreating scenes that might miss crucial distinctive features. Many image ReID models mistake similar-looking entities for each other, especially in Vehicle ReID.

Vehicle ReID is conceptually identical to person ReID, aiming to retrieve vehicles rather than people. Cars lack distinctive features and have similar background roads. In such cases, subtle details play a much more vital role in resolving similar vehicles. Prior vehicle ReID work includes handcrafted [23, 45] and deep learned features [4, 13, 21]. He et al. [9] is among the principal attempts to unify vehicle and person ReID. Image ReID methods are computationally expensive and infeasible for handling the temporal relations, making them unsuitable for videos.

Video person ReID: Common methodologies for generating video embeddings involve transformers [7], 3D convolutions [8, 14, 19], RNNs and LSTMs [2, 24, 25]. While the results are promising, these approaches often have memory constraints for a single GPU and involve many training parameters, making them infeasible for real-world surveillance systems. RNN based approaches often neglect intricate details within the frames, concentrating more on intra-frame structural ties. Gao and Nevatia [6] showed a temporal attention model could outperform LSTMs and 3D convolutions for ReID. Many 2D convolution-based methodologies have achieved significant ReID success, specifically those involving attention modules [5, 16, 17, 33].

Graph-based methods [1, 3, 37, 42] have been applied to both image and video person ReID. Shen et al. [32] treat images as graph nodes while disregarding the spatial subtleties. Yang et al. [43] use two branches for generating spatial and temporal relations, where nodes of graphs are segments of images. These approaches incur heavy computation and are ineffective at differentiating similar clothing styles, where the spatial structure is identical.

Fine-Grained Classification: Most fine-grained related research deals with images, either through enhancing image quality or recursively cropping critical regions and generating embeddings concurrently [11, 35, 46]. Such approaches fail on long videos with many frames. Generally, fine-grained ReID work has shown moderate success [29, 44, 55]. Recently, Zhang et al. [48] proposed multigranular attention for videos, surpassing the state-of-the-art.
Figure 4. The proposed video model has shown $t = 4$ video frames as input. Backbone CNN creates preliminary features, followed by a dimension reduction step. Global Feature Module creates a general overview of the entire clip. The fine-grained module highlights the spatial intricacies while the context module adds context via a non-local block. $\odot$ represent element-wise multiplication, $\text{softmax}$, and $\sum_c$ represent softmax, and summation along the channel dimension.

on MARS, iLIDS-VID, and PRID-2011 datasets. While such approaches are promising, they incur a high computational cost. Hence, this work builds upon Zhu et al. [56] using one-pass fine-grained rich embedding generation for videos allowing a computation-efficient implementation while adding context via non-local operations.

3. Methodology

Figure 4 shows our proposed model architecture. The proposed architecture has three major segments: a Global Feature Module (Section 3.2), a Fine-Grained Module (Section 3.3), and a Context Module (Section 3.4). The Global Feature Module averages feature spatially and temporally producing coarse-grained features ($\hat{f}_{\text{ImageNet}}$). The Fine-Grained Module creates spatial attention maps inspired by work on fine-grained image classification [56] in a parameter-less manner. The Context Module creates context-aware embeddings ($\hat{f}_{\text{MARS}}$) with the help of the fine-grained attention map. Concatenating $\hat{f}_{\text{ImageNet}}$ and $\hat{f}_{\text{MARS}}$ produces the final embeddings $f_{\ast}$. A shared weight classifier with softmax activation creates label vectors $Y_1$ and $Y_2$ corresponding to $\hat{f}_{\text{ImageNet}}$ and $\hat{f}_{\text{MARS}}$, respectively. The entire architecture employs only five $1 \times 1$ convolutions and two backbone ResNets, a single classifier, and two batch norm layers. We shall follow the $t \times h \times w \times c$ convention for the subsequent discussion to denote a tensor, indicating feature having $t$ frames ($t = 1$ for images) with $(h, w)$ spatial points and $c$ channels.

3.1. Backbone CNN Network

We use ResNet-50 as our backbone CNN for generating features for each video frame. We expand the receptive field of ResNet-50 by adjusting the last stride from $(2, 2)$ to $(1, 1)$ as described by Luo et al. [27]. Further, we follow the fine-grained image classification approach [56] of utilizing two ResNet-50(s): $\text{CNN}_{\text{ImageNet}}$ and $\text{CNN}_{\text{MARS}}$. $\text{CNN}_{\text{ImageNet}}$ is the generic ImageNet pre-trained CNN trying to capture the coarse-grained features of the image. Simultaneously, the other ResNet-50 ($\text{CNN}_{\text{MARS}}$) is the ReID task-specific CNN trying to capture fine-grain details; obtained by pre-training ResNet on large person ReID dataset (MARS dataset, produced from Pathak et al. [31]). Given a video clip of $t$ frames, $\text{CNN}_{\text{ImageNet}}$ and $\text{CNN}_{\text{MARS}}$ produce $f_{\text{ImageNet}}$ and $f_{\text{MARS}}$, respectively (both $\in \mathbb{R}^{t \times h \times w \times c}$). Further, two $1 \times 1$ convolution reduce dimension from $c$ to $c'$ ($c' = \frac{c}{2}$), producing $f'_{\text{ImageNet}}$ and $f'_{\text{MARS}}$.

3.2. Global Feature Module

Global Feature Module works on the $\text{CNN}_{\text{ImageNet}}$, capturing the overall foreground and background of the entire input video/image. The module uses global average pooling to average spatial features. The resulting feature vectors $A_{\text{gap}}$ are temporally averaged via temporal pooling (not needed for images). Finally, features are batch-normalized, producing coarse-grained features $\hat{f}_{\text{ImageNet}}$.

3.3. Fine-Grained Module

Fine-Grained Module produces fine-grained spatial attention maps, highlighting subtle intricacies within the spatial feature space. Traditionally, a weighted sum of channels learned via 2D convolution highlights these spatial regions. We argue against these pre-trained channel weights, which may not generalize well for unseen test classes. Hence, we deploy a run-time channel weighting technique based on the intuition that the significance of a particular channel is proportional to the average activation of its spatial feature map. Consequently, these weights are generated in run-time, making them ideal for ReID problems. We use $\text{softmax}$, and $\sum_c$ represent softmax, and summation along the channel dimension.
spatially averaged features $A_{gap}$ for weighting the channels.

$$S_{channel} = \text{softmax}(A_{gap}) \in \mathbb{R}^{t 	imes c^*}$$  \hspace{1cm} (1)

where $\text{softmax}$ is the softmax operation along the channel dimension. We apply these channel weights $S_{channel}$ to the absolute value of feature maps with sigmoid activation ($\sigma$):

$$f^+_{imageNet} = f_{imageNet} - \text{min}(f_{imageNet})$$  \hspace{1cm} (2)

$$A_{maps} = \sigma \left( \sum c^* f^+_{imageNet} \odot S_{channel} \right)$$  \hspace{1cm} (3)

where $A_{maps} \in \mathbb{R}^{t \times h \times w}$ is the spatial heat map, and $\text{min}$ returns the minimum value in the entire tensor. $\odot$ refers to pairwise multiplication of vectors. Such an approach is computationally more efficient.

### 3.4. Context Module

We use the context module via a non-local block [39] (also known as self-attention) on the task-specific features $f_{MARS}$ to introduce spatial and temporal context. The non-local block operation does a weighted sum of each point in spatial and temporal space ($THW$ space) to add context to the features. This approach is susceptible to noisy outliers, which may weaken the weight of significant regions. Hence, as an added safeguard, we shift our non-local block after the spatial attention to reduce the contribution of irrelevant regions.

$$A_1 = f_{MARS} \odot A_{map} \in \mathbb{R}^{t \times h \times w \times c^*}$$  \hspace{1cm} (4)

We pass the spatially attended features ($A_1$) through the non-local block, as shown in Figure 5. The traditional non-local block consists of a softmax-ed dot product between the query ($Q$) and key ($K$) vectors to assign weights to each key-value pair, followed by a weighted sum of value ($V$) vectors. The key-value pair act as the context for the query in $THW$ feature space (for images, its $HW$ feature space), giving more weight to contextually similar value vectors. We keep the query and key vectors same, saving around 0.497% (0.26 million) parameters.

$$Q = K = L_2^{norm} (\text{reshape}(\text{relu}(\theta_{1 \times 1}(A_1))))$$  \hspace{1cm} (5)

$$V = \text{reshape}(\text{relu}(\delta_{1 \times 1}(A_1)))$$  \hspace{1cm} (6)

where $\theta$ and $\delta$ are $1 \times 1$ convolution reducing channel dimension to $t \times \frac{r}{\tau}$. $L_2^{norm}$ does a $L_2$ normalization of $THW$ vectors. Matrix multiplication ($\odot$) between query $Q$ and key $K$ produces weight matrix $W$

$$W = \text{softmax}(Q^T \odot K) \in \mathbb{R}^{THW \times THW}$$  \hspace{1cm} (7)

$$V_{avg} = \text{reshape}(W \odot V) \in \mathbb{R}^{t \times h \times w \times c^*}$$  \hspace{1cm} (8)

We use equation 8 to do a weighted sum of each value vector $V$, followed by reshaping and $1 \times 1$ convolution ($\beta$) to restore input tensor dimension. Adding the resulting vector to the original $A_1$ yields contextually aware feature $A_2$:

$$A_2 = \beta_{1 \times 1}(V_{avg}) + A_1 \in \mathbb{R}^{t \times h \times w \times c^*}$$  \hspace{1cm} (9)

Additionally, we do weighted spatial averaging on $A_2$ via attentive pooling (equation 10):

$$A_3 = \frac{\sum_{(h,w)} A_2}{\sum_{(h,w)} A_{map}} \in \mathbb{R}^{t \times c^*}$$  \hspace{1cm} (10)

Final features $\hat{f}_{MARS}$ is obtained by passing $A_3$ through temporal pooling and batch normalization.

### 4. Experiments

#### 4.1. Loss Functions

We apply labeled smoothed cross-entropy loss exclusively on each label vector $Y_1$ and $Y_2$ and use their mean as the final classification loss $L^{CE}$. Among other loss functions, we apply batch hard triplet loss function ($L_{trip}$), center loss ($L_C$), and CL Centers OSM loss ($L_{OSM}$) [31, 40] on the final features $f^*$. Additionally, we apply a variant of variance regularization [12], which penalizes the same class feature ($f^*)^{c^2}$ variance across batch $B$, where $c^2$ indicates the unique set of identities in $B$.

$$R_{var} = \sum_{c \in B} \left( \frac{1}{\sum_i (f^*)^{c^2}} - \frac{1}{\sum_{i} (f^*)^{c^2}} \right)^2$$  \hspace{1cm} (11)

where $(f^*)^{c^2}$ indicates $i^{th}$ instance for the class $c^2$. Additionally, we apply KL divergence ($L_{cns}$) to align coarse-grained predictions $Y_1$ with the fine-grained ones $Y_2$:

$$L_{cns} = \max \left( \sum_{i} \frac{c}{\hat{C}} Y_1^i \log \left( \frac{Y_1^i}{Y_2^i} \right) - m_1, 0 \right)$$  \hspace{1cm} (12)

where $\hat{C}$ indicates the total number of classes, and $m_1$ is the margin parameter. We also use satisfied rank loss $L_{sr}$.
[56] to prevent $Y_1$ from dominating $Y_2$, consisting of an unbounded rank loss ($L_r$) and a limiting loss ($L_s$):

$$L_r = \max(0, Y_1^\pi - Y_2^\pi + m_2) \quad (13)$$

$$L_s = \min(L_s, L_r) \quad (15)$$

where $\pi \in \hat{C}$ indicates the correct class, $m_2$ and $m_3$ are margins parameters. While equation 13 prevents $Y_1$ from dominating $Y_2$, it may result in an unbounded solution. Equation 14 bounds the solution within limits. The total loss comprises a hyperparameter optimized\(^1\) weighted sum of all the losses mentioned above.

$$L_T = L_{CE}^{avg} + (1 - \beta) * L_{trip} + \beta * L_{OSM} + W_{var} * R_{var} + W_{cns} * L_{cns} + W_{sr} * L_{sr} \quad (16)$$

4.2. Datasets and Evaluation Metrics

Table 1 summarizes various datasets used for our experiments. While datasets like MARS, Market1501, and VeRi-776 use predefined splits, the other three datasets PRID-2011, iLIDS-VID, and CUHK01 are evaluated on random splits, averaged over ten iterations.

For evaluation, we followed the mean average precision (mAP) and the Cumulative Matching Characteristic for various ranks (rank-1 (R-1), rank-5 (R-5), rank-10 (R-10), and rank-20 (R-20)). We also perform evaluations using reranking (RR) [52]. For VeRi-776 and Market1501 datasets, we deploy spatial-temporal statistics (ST) [36] to boost our assessment precision.

4.3. Implementation Details

The input dimension for a single frame is $250 \times 150$ with $t = 4$ frames on the MARS and PRID-2011 datasets. We set the batch size $B = 32$ and the number of positive instances per batch $K = 5$. The iLIDS-VID dataset has input size of $220 \times 150$ with $t = 5$, $B = 28$, and $K = 5$. For all image person ReID datasets, we set the input size to $250 \times 150$, with $B = 128$ and $K = 4$. For the VeRi-776 dataset, the input size is $150 \times 250$, with the task-specific backbone CNN, pre-trained on the VehicleID dataset [21]. During the training, we include techniques proposed by Luo et al. [27], namely warm-up learning rate and random erasing [53]. For evaluation, we average an input frame and its horizontally flipped mirror image embedding, followed by L2 normalization. For videos, we normalize the entire video embedding. Our unified approach for image and video ReID aids us in image ReID by pre-training FGReID for video dataset MARS, denoted by FGReID*.

4.4. Comparison with State-of-the-art Methods

We compare FGReID with existing state-of-the-art (SOTA) using datasets in Table 1. We report re-rank (RR) accuracy separately to maintain uniformity with previous works. Robust mAP, R-1, R-5, and R-20 accuracy highlights the importance of fine-grained details and context information for generating robust embeddings for videos and images. The success of previous work involving temporal self-attention (GLTR [15]) and multi-granular attentive features [48] supports our use of non-local block (self-attention) and fine-grained spatial attention.

4.4.1 Video ReID

Zhang et al. [48] proposed aggregating spatial and temporal features by splitting channels into $S$ groups (SG-RAFA, $S = 1$) and reading fine-grained spatial details on $N$ granular scales (MG-RAFA, $N = 2, 4$).

Table 2 shows FGReID surpassing SOTA by 0.3% on mAP and 0.2% on R-20 accuracy without re-rank (RR) on the MARS dataset. With re-rank, FGReID beats SOTA by 1.1%, 0.8% on R-1, 0.6% on R-5 and 0.3% on R-20 accuracy, whereas on iLIDS-VID, FGReID outperforms existing SOTA significantly, with a margin of 2.9% on R-1, and 1.2% on R-5 accuracy, with 100% accuracy on R-20 (Ta-
Table 2. Performance evaluation on MARS dataset.

| Method                  | mAP | R-1  | R-5  | R-20 |
|-------------------------|-----|------|------|------|
| w/o RR                  |     |      |      |      |
| MGH [42]                | 85.8| 90.0 | 96.7 | 98.5 |
| SG-RAFA [48]            | 85.1| 87.8 | 96.1 | 98.6 |
| MG-RAFA(N=2) [48]       | 85.5| 88.4 | 97.1 | 98.5 |
| MG-RAFA(N=4) [48]       | 85.9| 88.8 | 97.0 | 98.5 |
| FGReID                  | 86.2| 89.6 | 97.0 | 98.8 |

with RR

| Pathak et al. [31]      | 88.5| 88.0 | 96.1 | 98.5 |
| FGReID                  | 89.6| 88.8 | 96.7 | 98.8 |

Table 3. Results comparison for the iLIDS-VID dataset. FGReID achieves 99.8% for R-10 accuracy.

| Method                  | R-1  | R-5  | R-10 | R-20 |
|-------------------------|------|------|------|------|
| GLTR [15]               | 86.0 | 98.0 | -    | -    |
| Zhao et al. [49]        | 86.3 | 97.4 | 99.7 | -    |
| MG-RAFA (N=4) [48]      | 88.6 | 98.0 | 99.7 | -    |
| FGReID                  | 91.5 | 99.2 | 100  | -    |

Table 4. Results comparison for the PRID-2011 dataset.

| Method                  | mAP | R-1  | R-5  |
|-------------------------|-----|------|------|
| w/o (RR)                |     |      |      |
| st-ReID (ST) [36]       | 87.6| 98.1 | 99.3 |
| Adaptive L2 Reg [30]    | 88.9| 95.6 | -    |
| FGReID                  | 86.1| 94.0 | 97.6 |
| FGReID*                 | 87.1| 94.7 | 98.5 |
| FGReID* + ST            | 91.0| 98.2 | 99.4 |

with RR

| st-ReID (ST) [36]       | 95.5| 98.0 | 99.9 |
| st-ReID (ST) + UnityStyle [20] | 95.8| 98.5 | 99.0 |
| FGReID* + ST            | 96.0| 98.1 | 99.0 |

Table 5. Results comparison on Market1501 dataset. R-10 (and RR) accuracy for FGReID* + ST is 99.4% (99.3%).

Table 6. Results for CUHK01 dataset. Accuracy of FGReID* for p=486 (R-10) is 98.7% and p=100 (R-5) is 99.8%.

| Method                  | p = 486 | p = 100 |
|-------------------------|---------|---------|
|                         | R-1     | R-5     | R-10    | R-5    |
| GLTR [15]               | 90.4    | 97.8    | -       | -      |
| BraidNet [41]           | -       | 93.04   | 99.97   |        |
| FGReID                  | 89.6    | 96.7    | 98.9    | 99.8   |
| FGReID*                 | 90.9    | 97.5    | 99.1    | 99.8   |

Table 7. Results comparison on the VeRi-776 dataset. R-10 (and RR) accuracy for the FGReID + ST is 99.5% (98.3%).

Table 8. All times are in seconds, while M means parameters count in millions.

| Method                  | Size  | Train Time | Eval Time |
|-------------------------|-------|------------|-----------|
| MGH [42]                | 44.18M| 414.90     | 719.87    |
| Pathak et al. [31]      | 91.90M| 71.03      | 264.95    |
| GLTR [15]               | 24.77M| 62.08      | 225.43    |
| FGReID                  | 52.64M| 130.82     | 458.28    |

4.4.2 Image ReID

The original st-ReID (SOTA) [36] uses a simple part-based convolutional model with spatial-temporal statistics (ST). FGReID pre-trained on MARS dataset (with ST) surpasses SOTA on the Market1501 dataset by 3.6% on mAP, 0.2% on R-1, and 0.1% on R-5 accuracy, w/o RR. With RR, FGReID surpasses SOTA by 0.2% on mAP accuracy. All the other competing SOTAs are shown in Table 5. For CUHK01, Table 6 shows the performance of FGReID across both the splits, p=486, and p=100. Table 7 shows FGReID exceeding PRN (SOTA) [4] on R-1 accuracy by 1.3% (w/o RR) on the VeRi-776 dataset.

4.5. Memory and Computation Comparison

We compare the publicly available SOTAs w.r.t. to their memory sizes, train time (forward and backward passes), and evaluation time (forward pass) under identical conditions on MARS datasets with identical loss functions on a p40 GPU. For the train time, we train the model for 100 epochs while averaging the time for the last 50 epochs. Similarly, we repeat evaluations (eval) ten times, averaging the time for the last five runs. Table 8 shows FGReID stands third in rank w.r.t. memory constraints and training speed, although it shows far superior performance than the above two rank holders. GLTR [15] requires the least computation and a small number of training parameters owing to a simple
convolution and a non-local block operation. The method by Pathak et al. [31] has the highest number of parameters because of the temporal attention architecture, while simple spatial averaging contributes to a shorter training time. Top-performing models are multi-granular approaches (MGH [42], MG-RAFA [48]), which do repeated calculations for multi-granular scales to capture fine-grained details, giving them slow training and evaluation times.

For future works, keeping the $1 \times 1$ conv and batch normalization layer same for the coarse-grained and fine-grained branches would save around 2.36 million parameters. Finding a substitution for the fine-grained ResNet with a lightweight backbone CNN would reduce in excess of 23.51 million parameters.

5. Ablation Study

5.1. Visualization

Case against Mistaken Identity: We manually searched the MARS dataset for similar-looking identities to show the effectiveness of fine-grained details in tackling mistaken identity (similar-looking individuals). We compare the dot product score of the query tracklet and a corresponding correct tracklet (R-1) to the dot product score of the query tracklet and a wrong similar-looking tracklet. Figure 6 shows the models accurately differentiating between similar looking identities by giving a high score to the correct class. In Figure 6, scores suggest that the model pays attention to minute details like backpack design ((a), (b), (e)) orange handbags ((f)), etc., to re-identify individuals. The results also indicate that the model is invariant to the color structures of the frames ((c), (d), (g)), reducing the chances of mistaking identities.

Visualization of Fine-grained Attention: Our model produces more run-time oriented spatial attention maps without incurring extra training parameters in contrast to traditional convolutional attention mechanisms. Figure 7 shows our model spotlights fine-grained features like handbags and shoes while disregarding backgrounds and other non-distinctive details. These fine-grained regions add context into the frames, zeroing out regions in spatial and temporal space which are missing these subtleties.

5.2. Analysis of hyperparameters

We examine the role of various hyperparameters and loss functions in our methodology. Unless explicitly mentioned, we take $B = 128$ and $K = 4$ for the CUHK01 dataset ($p=486$, split=0). For iLIDS-VID dataset we take $B = 32$, $K = 4$ and $t = 4$. Input frame/image sizes are $224 \times 112$.

Batch size We conduct a series of experiments on the CUHK01 and iLIDS-VID datasets to determine the optimal $B$ and $K$ for the image and video datasets. Figure 8(a) shows optimal performance at $K = 2, 4, 5$ for $B = 80$ on the CUHK01 dataset. Similarly, Figure 8(b) shows peak performance at $K = 4$ with $B = 32$ for the iLIDS-VID dataset. Figure 8(c) shows the impact of different ratios of negative instances to positive instances per batch $\gamma$, for constant $K = 4, 5$. For $K = 4$ and $K = 5$ the optimal ratio $\gamma$ is around 27:1 ($B = 112$) and 23:1 ($B = 120$) respectively. For the iLIDS-VID dataset, optimal performance is around $B = 28$ for $K = 5$ (Figure 8(d)), approximating the $\gamma$ to 23:5. Compared to videos, $\gamma$ play a more significant role in images.
Figure 8. (a) and (b) evaluates different K for constant B. (c) and (d) varies B while keeping constant K.

Figure 9. (a) compares performance for the different number of video frames (t). (b) shows the performance impact of various input image dimensions.

**Number of Video frames**: For our video model, we have an additional hyperparameter for video length t, indicating the number of frames used for ReID. Figure 9(a) shows t = 3, 4, 5 performing optimally for the iLIDS-VID dataset.

**Image/Frame Size**: Most ReID works adhere to a conventional height H to width W ratio of 2:1, predominantly having shapes 224 × 112 and 256 × 128. We argue such a ratio may distort human proportions with Figure 9(b) showing the optimal performance with W = 150 and H = 220, 250. Missing regions are out-of-memory locales.

### 5.3. Analysis of Loss functions

To study the significance of various loss functions, we exclude one loss from the total loss and assess the best performing model. As shown in Table 9, the absence of classification (L_{avg}^{CE}) and KL divergence (L_{cns}) losses have a significant unfavorable effect, indicating a strong need for alignment of coarse-grained and fine-grained predictions. We also observed the absence of center loss (L_{C}) or satisfied rank loss (L_{sr}) slowed the convergence significantly.

### 6. Ethical Consideration

Our approach intends to create fine-grained rich embeddings for videos and images in a zero-shot learning setting, generalizing to various embedding related tasks. Once properly deployed, ReID can spare hours of human effort in tracing suspects by reducing city-wide camera footage to a minimal subset. But possible unintended use cases exist, including the unapproved tracking of individuals and targeting protesters. The authors have genuine concerns over the alleged targeting of Uighur Muslims in China using ReID. This unintentional application is undesirable, and to reduce the likelihood of it happening in the future, we have chosen not to release any hyperparameters or trained weights. All research into re-identification should consider and curtail these potentials for misuse.

### 7. Conclusion

This paper proposes FGReID, a lightweight method for generating contextually coherent and fine-grained rich embeddings for ReID tasks. FGReID can handle similar-looking identities and offers a unified solution for both video and image ReID tasks. The pipeline consists of three vital components: a global feature module, a fine-grained module, and a context module. The global feature branch delivers the general overview, while the fine-grained module highlights the minute subtleties. The context module adds temporal and spatial context via a non-local block operation to the spatially attended features. Several experiments show the viability of FGReID for both video and image ReID tasks. We perform an ablation study to show the significance of various training hyperparameters.

### 8. Acknowledgement

This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise. Mr. Shenglong Wang (NYU), Dr. Amir Erfan Esfratifar (USC), Mr. Parth McPherson (NYU), and Dr. Thomas Lux (Virginia Tech) provided their invaluable contributions to the entire project.

---

Table 9. Impact of various loss functions on the performance of models. Baseline is trained on all loss functions.
References

[1] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3356–3365, 2017. 2

[2] Xiang Bai, Mingkun Yang, Tengteng Huang, Zhiyong Dou, Rui Yu, and Yongchao Xu. Deep-person: Learning discriminative deep features for person re-identification. Pattern Recognition, 98:107036, 2020. 1, 2

[3] L. Bao, B. Ma, H. Chang, and X. Chen. Preserving structural relationships for person re-identification. In 2019 IEEE International Conference on Multimedia Expo Workshops (ICMEW), pages 120–125, 2019. 1, 2

[4] Hao Chen, Benoit Lagadec, and Francois Bremond. Partition and reunion: A two-branch neural network for vehicle re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019. 1, 2, 6

[5] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas S. Huang. STA: spatial-temporal attention for large-scale video-based person re-identification. CoRR, abs/1811.04129, 2018. 2

[6] Jiyang Gao and Ram Nevatia. Revisiting temporal modeling for video-based person reid. arXiv preprint arXiv:1805.02104, 2018. 1, 2

[7] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video Action Transformer Network. In CVPR, 2019. 2

[8] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In ECCV, 2020. 2

[9] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification, 2020. 2

[10] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person Re-Identification by Descriptive and Discriminative Classification. In Proc. Scandanavian Conference on Image Analysis (SCIA), 2011. 5

[11] Tao Hu and Honggang Qi. See better before looking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019. 1, 2

[12] Ranran Huang, Hanbo Sun, Ji Liu, Lu Tian, Li Wang, Yi Shan, and Yu Wang. Feature variance regularization: A simple way to improve the generalizability of neural networks, 2019. 2, 4

[13] Tsung-Wei Huang, Jianui Cai, Hao Yang, Hung-Min Hsu, and Jenq-Neng Hwang. Multi-view vehicle re-identification using temporal attention model and metadata re-ranking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019. 1, 2

[14] 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. 2

[15] J. Li, S. Zhang, J. Wang, W. Gao, and Q. Tian. Global-local temporal representations for video person re-identification. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 3957–3966, 2019. 1, 5, 6

[16] Mengliu Li, Han Xu, Jinjun Wang, Wenheng 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 (WACV), March 2020. 1, 2

[17] S. Li, S. Bak, P. Carr, and X. Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 369–378, 2018. 2

[18] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In ACCV, 2012. 5

[19] Xingyu Liao, Lingxiao He, Zhouwang Yang, and Chi Zhang. Video-based person re-identification via 3d convolutional networks and non-local attention. In C.V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, Computer Vision – ACCV 2018, pages 620–634, Cham, 2019. Springer International Publishing. 2

[20] C. Liu, X. Chang, and Y. D. Shen. Unity style transfer for person re-identification. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6886–6895, 2020. 2, 6

[21] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang. Deep relative distance learning: Tell the difference between similar vehicles. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2167–2175, 2016. 2, 5

[22] X. Liu, W. Liu, H. Ma, and H. Fu. Large-scale vehicle re-identification in urban surveillance videos. In 2016 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6, 2016. 5

[23] Xinchen Liu, Huadong Ma, Huiyuan Fu, and Mo Zhou. Vehicle retrieval and trajectory inference in urban traffic surveillance scene. In Proceedings of the International Conference on Distributed Smart Cameras, ICDSC ’14, New York, NY, USA, 2014. Association for Computing Machinery. 2

[24] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):8786–8793, Jul. 2019. 2

[25] Z. Liu, T. Chen, E. Ding, Y. Liu, and W. Yu. Attention-based convolutional lstm for describing video. IEEE Access, 8:133713–133724, 2020. 2

[26] Chuanchen Luo, Yuntao Chen, Naiyuan Wang, and Zhaoxiang Zhang. Spectral feature transformation for person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019. 1, 2

[27] Hao Luo, Youzhi Gu, Xingyu Liao, S. Lai, and W. Jiang. Bag of tricks and a strong baseline for deep person re-identification. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1487–1495, 2019. 2, 3, 5

[28] H. Luo, W. Jiang, X. Fan, and C. Zhang. Sinfred: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. IEEE Transactions on Multimedia, pages 1–1, 2020. 2
[29] Dechao Meng, Liang Li, Shuhui Wang, Xingyu Gao, Zheng-Jun Zha, and Qingming Huang. Fine-grained feature alignment with part perspective transformation for vehicle reid. In Proceedings of the 28th ACM International Conference on Multimedia, MM ’20, page 619–627, New York, NY, USA, 2020. Association for Computing Machinery.

[30] Xinyang Ni, Liang Fang, and Heikki Huttunen. Adaptivevid: Adaptive I2 regularization in person re-identification. arXiv preprint arXiv:2007.07875, 2020.

[31] Priyank Pathak, Amir Erfan Eshratifar, and Michael Gormish. Video person re-id: Fantastic techniques and where to find them. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), April 2019.

[32] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaohua Xie. Spatial-temporal person re-identification. IEEE Access, 7:22457–22470, 2019.

[33] 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 on Computer Vision (ECCV), September 2018.

[34] A. Tan, G. Zhou, and M. He. Rapid fine-grained classification of butterflies based on fcm-km and mask r-cnn fusion. IEEE Access, 8:124722–124733, 2020.

[35] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiao-hua Xie. Spatial-temporal person re-identification. Proceedings of the AAAI Conference on Artificial Intelligence, pages 8933–8940, 2019.

[36] Guan’an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[37] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, Computer Vision – ECCV 2014, pages 688–703, Cham, 2014. Springer International Publishing.

[38] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[39] Xinhao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, and Neil Martin Robertson. Deep metric learning by online soft mining and class-aware attention. CoRR, abs/1811.01459, 2018.

[40] Y. Wang, Z. Chen, F. Wu, and G. Wang. Person re-identification with cascaded pairwise convolutions. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1470–1478, 2018.

[41] 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 (CVPR), June 2020.

[42] Jinrui Yang, Wei-Shi Zheng, Qiye 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 (CVPR), June 2020.

[43] Zhao Yang, Zhigang Chang, and Shibao Zheng. Large-scale video-based person re-identification via non-local attention and feature erosion. In Guangtao Zhai, Jun Zhou, Hua Yang, Ping An, and Xiaokang Yang, editors, Digital TV and Wireless Multimedia Communication, pages 327–339, Singapore, 2020. Springer Singapore.

[44] D. Zapletal and A. Herout. Vehicle re-identification for automatic video traffic surveillance. In 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1568–1574, 2016.

[45] L. Zhang, S. Huang, W. Liu, and D. Tao. Learning a mixture of granularity-specific experts for fine-grained categorization. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 8330–8339, 2019.

[46] Z. Zhang, C. Lan, W. Zeng, and Z. Chen. Densely semantically aligned person re-identification. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 667–676, 2019.

[47] Z. Zhang, C. Lan, W. Zeng, and Z. Chen. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10404–10413, 2020.

[48] Yi Zhao, X. 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 (CVPR), June 2019.

[49] 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 Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, Computer Vision – ECCV 2016, pages 868–884, Cham, 2016. Springer International Publishing.

[50] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, and Qi Tian. 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 (CVPR), June 2019.

[51] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, and Qi Tian. Scalable person re-identification: A benchmark. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV ’15, page 1116–1124, USA, 2015. IEEE Computer Society.

[52] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. CoRR, abs/1807.09975, 2018.

[53] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), April 2019.

[54] J. Zhou, B. Su, and Y. Wu. Online joint multi-metric adaptation from frequent sharing-subset mining for person re-
identification. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2906–2915, 2020. 2

[55] Q. Zhou, B. Zhong, X. Lan, G. Sun, Y. Zhang, B. Zhang, and R. Ji. Fine-grained spatial alignment model for person re-identification with focal triplet loss. *IEEE Transactions on Image Processing*, 29:7578–7589, 2020. 2

[56] Youxiang Zhu, Ruochen Li, Yin Yang, and Ning Ye. Learning cascade attention for fine-grained image classification. *Neural Networks*, 122:174–182, 2020. 1, 3, 5