

# Attribute-Driven Feature Disentangling and Temporal Aggregation for Video Person Re-Identification

Yiru Zhao<sup>1,2\*</sup>, Xu Shen<sup>2</sup>, Zhongming Jin<sup>2</sup>, Hongtao Lu<sup>1†</sup>, Xian-sheng Hua<sup>2‡</sup>

<sup>1</sup> Key Lab of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering,  
Department of Computer Science and Engineering,  
MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University  
<sup>2</sup>Alibaba Damo Academy, Alibaba Group

{yiru.zhao,htlu}@sjtu.edu.cn, {shenxu.sx,zhongming.jinzm,xiansheng.hxs}@alibaba-inc.com

## Abstract

*Video-based person re-identification plays an important role in surveillance video analysis, expanding image-based methods by learning features of multiple frames. Most existing methods fuse features by temporal average-pooling, without exploring the different frame weights caused by various viewpoints, poses, and occlusions. In this paper, we propose an attribute-driven method for feature disentangling and frame re-weighting. The features of single frames are disentangled into groups of sub-features, each corresponds to specific semantic attributes. The sub-features are re-weighted by the confidence of attribute recognition and then aggregated at the temporal dimension as the final representation. By means of this strategy, the most informative regions of each frame are enhanced and contributes to a more discriminative sequence representation. Extensive ablation studies verify the effectiveness of feature disentangling as well as temporal re-weighting. The experimental results on the iLIDS-VID, PRID-2011 and MARS datasets demonstrate that our proposed method outperforms existing state-of-the-art approaches.*

## 1. Introduction

Person re-identification (Re-ID) is at the core of intelligent video surveillance systems because of a wide range of potential applications. Given a query person, the task aims at matching the same person from multiple non-overlapping cameras. It remains a very challenging task due to the large variations of human poses, occlusions, viewpoints, illuminations and background clutter.

\*This work was done when the author was visiting Alibaba as a research intern.

†Corresponding author.

‡Corresponding author.

Image-based single-query re-id task has been widely investigated in recent years, including feature representation [15, 21, 44] and distance metric learning [19, 38, 27]. Deep learning methods have shown significant advantages in feature learning and have been proven highly effective in person re-id tasks [18, 5, 32, 35, 30]. Existing works have shown that multi-query strategy obviously outperforms single-query by simply pooling features across a track-let [43, 48, 13]. This improvement is almost cost-free because multi-frame context is easily available by visual tracking in real-world surveillance applications.

The video informations are further explored to extract temporal features, nourishing a series of video-based re-id approaches. Some works [26, 40, 4] involve optical flow to provide motion features. Recurrent neural networks are applied in [49, 40, 4] to explore the temporal structure of input image sequences. Temporal Attention models are also utilized in [49, 40, 17] to replace temporal average pooling, motivated by the assumption that the frames with higher quality and less occlusions ought to have larger weights in aggregation. Local features of body regions have been used in previous works [43, 42, 39] and have shown superior for fine-grained identification. While in the video-based re-id task, it is suboptimal for local features of the same body region from different frames to share equal temporal weights due to the various human poses and occlusions within the image sequences. Our proposed method is motivated mainly by this observation and is designed to enhance the more informative frames of each regions.

An example of our proposed method is shown in Fig. 1. The feature of one frame is disentangled into several sub-features corresponding to specific semantic attribute groups. In the displayed image sequences, frame-1 captured clear frontal face so it has higher weight in *Head* group. While the bag is invisible in frame-1, the weights of *Bag* groups are mainly concentrated on frame-2 and frame-3. Frame-2

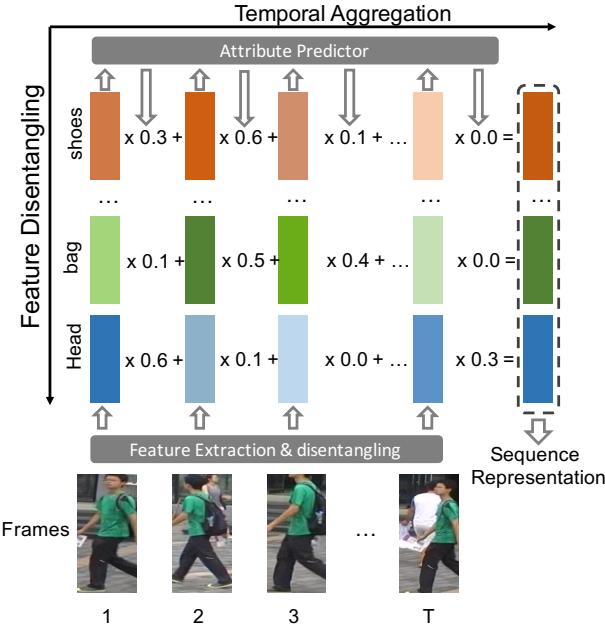


Figure 1. Illustration of our approach. The feature of one frame is disentangled into several sub-features corresponding to a specific semantic group. In each group, the sub-features from  $T$  frames are aggregated with adaptive weights. The aggregated sub-features are concatenated as the final feature representation of this sequence

also has the highest weight in *Shoes* group. The weights of frame-T are relatively low because of the poor detection bounding box and clutter background. The re-weighted sub-features are aggregated at the temporal dimension and then concatenated as the representation of the input sequence. We refine the temporal weights to the sub-feature level for handling various poses, occlusions and detection localizations within the sequence.

It is worth nothing that our proposed method relied on attribute annotations. However, it is labor expensive to manually annotate attribute labels for each identity in real-world applications. To address this problem, we introduce a transfer learning algorithm to automatically annotate attribute labels on re-id dataset by utilizing the knowledges learned on the attribute dataset.

## 2. Related Work

Related works of the proposed method can be summarized into three categories: image-based person re-id, video-based person re-id and attribute learning. We will explain the connections and differences between our work and these methods in the corresponding aspects.

### 2.1. Image-based person re-id

Person re-id is a challenging task which has been investigated for several years, while it still faces the same problems of various viewpoints, poses, illuminations and occlusions as other computer vision problems. Previous works mainly develop their solutions from two categories: extracting reliable feature representations [15, 21, 44] and learning robust distance metrics [19, 38, 27]. By means of the development of convolutional neural networks(CNNs), a lot of recent re-id models are designed based on CNN structure [1, 8, 6, 18, 31, 34, 42]. For example, [1] propose a method for simultaneously learning features and a corresponding pairwise similarity metric for person re-id. [8] present a scalable deep feature learning model for person re-id via relative distance comparison.

The vanilla CNN models only produce global features, while local details of body regions have been proven effective in person re-id task [42, 43, 39]. [42] propose a method which learns features of different body regions by a multi-stage ROI pooling network. [43] present a part-aligned representation approach to handle the body misalignment problem with attention model. [39] propose an attention-aware network to deal with the misalignment and occlusion problem by human parsing. In our proposed method, person attributes are utilized for learning local details and disentangling features into semantic groups, which also align the sub-feature for temporal fusion.

### 2.2. Video-based person re-id

Image-based re-id can be naturally extended to multi-shot re-id in real-world applications with the track-lets detected in video sequences. Recent works begin to explore video-based re-id problem. [26, 40, 4] involve optical flow calculated between adjacent frames as the input data, which provide motion features such as gait pattern. However, the calculation of optical flow is time-consuming, which is impractical in real-time applications. [49, 40, 4] apply recurrent neural networks (RNNs) on sequence of single-shot features to explore the temporal structure. Average pooling is a common strategy to merge features at the temporal dimension, while [49, 40, 17] utilize attention model to selectively focus on the most informative frames. In order to maximize the discriminability of each person regions, we further refine the temporal weights from feature level to sub-feature level in our method. [40, 4, 41] design siamese networks which take pairs of sequences as input and verify whether they belong to the same identity. The siamese architecture improves performance by pairwise comparison but is time-consuming in large-scale retrieval. On the contrary, our single-pass method only extracts features on each sequence once, which is efficient for real-time applications.

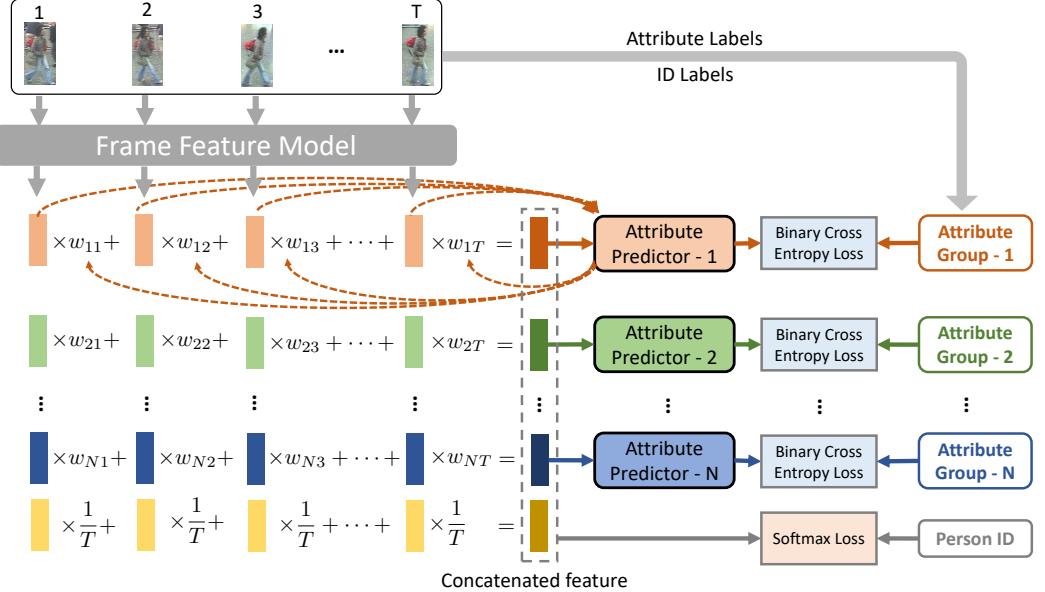


Figure 2. Architecture of our method. The attribute labels are split into  $N$  groups. Frame features are disentangled into  $N + 1$  segments,  $N$  of which correspond to the  $N$  attribute groups and one for global representation. The temporal weights  $w_{nt}$  are calculated by the recognition confidence, which do not provide gradients for training stability, as shown in dash arrow. Attribute predictor in each group is trained on the merge sub-features. The concatenated feature, including  $N + 1$  merge sub-features, represents the input sequence.

### 2.3. Attribute learning

Attribute learning [3, 2, 23] has attracted much attention in face identification [33, 37] as well as person re-id [31, 20, 29]. Previous works prove that the discriminability of recognition model can be improved by correctly predicting attributes. [37] propose a joint deep architecture for face recognition and facial attribute prediction. [31] address the person re-id problem with attributes triplet loss and improve the performance. [20] demonstrate that re-id task benefit from the multi-task learning process.

Different from existing multi-task methods that simply add an attribute prediction loss, our method utilizes attributes to disentangle features into semantic groups and further calculate the temporal weights of each sub-features. The annotation cost of attribute labels limits the expansion of attribute based methods in real-world scenarios. To address this problem, our method obtains attribute labels by transfer learning, without additional annotation cost.

## 3. Proposed Method

### 3.1. Feature Disentangling and Temporal Aggregation

In this section, we will introduce how to produce the feature of an input sequence with the attribute labels, and the model architecture is shown in Fig. 2.

**Frame Sampling.** The sequence lengths in video re-id task usually vary greatly, and a common practice is to sam-

ple a sequence of fixed frame number  $T$ . Existing RNN-based approaches require continuous frames as the input. However, a short segment of continuous video frames are highly correlated and are not much more informative than single image. On the contrary, the entire video often contains variant visual appearances (*e.g.* viewpoints, body poses). In order to utilize visual information from the entire video, we equally divide the sequence into  $T$  chunks  $\{C_t\}_{t=1}^T$ . One frame  $f_t$  is randomly sampled from each chunk  $C_t$ , then the entire video is represented by the set of sampled frames  $\{f_t\}_{t=1}^T$ .

**Feature Disentangling.** The next step is to produce the sequence feature with the sampled frames. Due to the various human poses and occlusions in the sequence, the informative local regions of each frame ought to be enhanced. Hence we firstly disentangle the frame feature into several groups and then calculate the temporal weights for each sub-features.

We adopt ResNet [12] for feature extraction. The global feature, *i.e.* a full-connected layer  $fc_1$  after avg-pooling of *Residual\_Block\_4*, is split into  $N + 1$  segments,  $N$  of which correspond to  $N$  local attribute groups and one for global representation.

$$fc_1 \rightarrow [fc^1, \dots, fc^N, fc^{N+1}] \quad (1)$$

According to the attributes in RAP dataset [16], we set  $N = 6$  in our method and the attribute groups are listed in Table 1. Each sub-feature is associated with an attribute

Table 1. Semantic attribute groups used in our method. Example attributes of each group are also listed below.

Group	Attributes
Gender & Age	Female, AgeLess16, ..., Age31-45
Head-Shoulder	Hat, Glasses, ..., BlackHair
Up-Body	Shirt, SuitUp, ..., up-Blue
Low-Body	Dress, Skirt, ..., low-Black
Shoes	Sport, Leather, ..., shoes-White
Attach	Backpack, HandBag, ..., PlasticBag

group by an attribute predictor  $AP_n$ , which consists of a fully-connected layer and a sigmoid layer to predict all the binary attributes in the  $n$ -th group. Driven by the attribute prediction loss, the global features are disentangled to represent  $N$  groups of local regions and the sub-features of each frame are aligned.

**Temporal Aggregation.** Next, we need to merge  $T$  sub-features from the sampled sequence at the temporal dimension. A common practice is average pooling, *i.e.* all sub-feature has the same weight  $1/T$ . However, not all frames are equally informative due to the variations of human poses, occlusions and viewpoints. We are more concerned about the frames which provide explicit attribute information, so we calculate the weight  $w_{nt}$  of the  $t$ -th frame in the  $n$ -th group by the attribute recognition confidence. Specifically, the confidence is calculated by the entropy of the attribute prediction score:

$$Conf(p) = e^{\frac{Ent(p)}{\sigma^2}}, Ent(p) = \frac{1}{A_n} \sum_{i=1}^{A_n} p_i \log(p_i) \quad (2)$$

where  $A_n$  is the number of attributes in the  $n$ -th group,  $p_i$  is the prediction result of the  $i$ -th attribute in the group,  $\sigma$  is a hyper-parameter to control the degree of re-weighting. Then the confidence scores of  $T$  frames are normalized to obtain the temporal weights:

$$w_{nt} = \frac{Conf(AP_n(fc_t^n))}{\sum_{i=1}^T Conf(AP_n(fc_i^n))} \quad (3)$$

Then the sub-features are aggregated with the temporal weights to the merged representation:

$$fc_{merge}^n = \sum_{t=1}^T w_{nt} fc_t^n \quad (4)$$

$fc_{merge}^n$  is finally utilized to train the attribute predictor  $AP_n$  by Binary Cross Entropy loss with the attribute labels. It is worth noting that the calculation of temporal weights  $w_{nt}$  does not contribute to the back propagation for the training stability, as denoted by the dashed line in Fig. 2.

Besides the  $N$  sub-features for local regions, the global sub-features are merged with equal weights  $1/T$ . Finally,

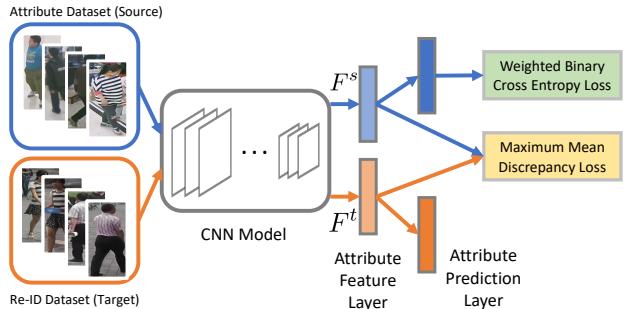


Figure 3. Illustration of the attribute transfer learning model, which learns to recognize attributes by optimizing the Weighted Binary Cross Entropy loss. The Maximum Mean Discrepancy loss is utilized to regularize the feature distribution between source and target domain.

the entire input sequence is represented by concatenating the  $N + 1$  merged sub-features  $[fc_{merge}^1, \dots, fc_{merge}^{N+1}]$ . At the training stage, softmax loss on the concatenated feature and  $N$  attribute prediction losses on the merged sub-features are deployed to train the whole network. At the testing stage, the similarity of video sequences is evaluated by Euclidean distance of concatenated features after  $L_2$ -normalization.

### 3.2. Transfer Learning for Attribute Recognition

Our proposed method relies on attribute labels for feature disentangling and temporal aggregation. Different from existing works [20] which require expensive labor to manually annotate attribute labels on person re-id datasets, we transfer attribute information from person attribute datasets to re-id datasets. By means of transfer learning, no additional annotation cost are required so that this method can be easily extended to other datasets and more scenarios.

Given a person attribute dataset (source domain), a direct practice for generating attribute labels on re-id dataset (target domain) is training an attribute recognition model first and then predict labels on the re-id images. However, the attribute recognition model trained only with source set is suboptimal on the target set due to the non-ignorable domain gap. The inconsistent feature distributions influence the attribute prediction on the re-id dataset.

Under the assumption that the person images (both in source and target datasets) share the same set of semantic attributes, the distribution distance of the attribute feature space between the source set and the target set ought to be minimize. The architecture is shown in Fig. 3 and a CNN model is designed to recognize person attributes. The penultimate layer is the attribute feature layer (denoted by  $F$ ) and the last layer is the prediction layer. We use the Maximum Mean Discrepancy (MMD) [11, 24, 25] to mea-

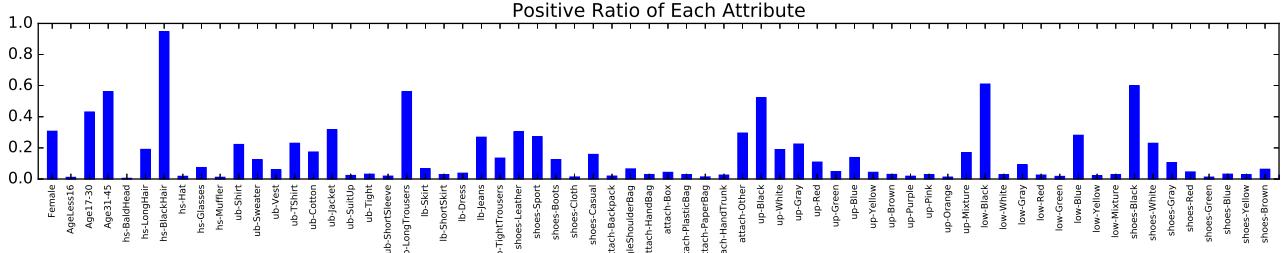


Figure 4. Positive ratios of the selected attributes on RAP dataset. Many attributes are extremely unbalanced.

sure the distance between two distribution. Given the source and target attribute features  $\{F^s\}_i^{n_s}$ ,  $\{F^t\}_i^{n_t}$  in each mini-batch, the MMD loss can be calculated by:

$$\mathcal{L}_{MMD} = \frac{1}{n_s^2} \sum_i \sum_j k(F_i^s, F_j^s) + \frac{1}{n_t^2} \sum_i \sum_j k(F_i^t, F_j^t) - \frac{2}{n_s n_t} \sum_i \sum_j k(F_i^s, F_j^t) \quad (5)$$

We select the Gaussian kernel with  $\alpha = 0.5$  as the kernel function  $k$ :

$$k(F_i^s, F_j^t) = \exp\left(-\frac{\|F_i^s - F_j^t\|^2}{2\alpha^2}\right) \quad (6)$$

The distribution variance of attribute feature space between the attribute dataset and re-id dataset is regularized by the MMD loss  $\mathcal{L}_{MMD}$ .

The attribute feature layer is followed by a fully-connected layer for attribute recognition. The outputs are activated by Sigmoid to predict the binary attributes. A widely used loss for binary label is Binary Cross Entropy (BCE) loss:

$$\mathcal{L}_{BCE} = -\frac{1}{L} \sum_{i=1}^L y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \quad (7)$$

where  $L$  is the number of attributes,  $p_i$  is prediction probability of the  $i$ -th attribute and  $y_i$  is the corresponding label.

However, most of the binary attributes are unbalanced, as shown in Fig. 4. The positive ratios of rare attributes (e.g. *ub-SuitUp*, *low-Red*) are quite small. The model trained with BCE loss prefer to output common attributes due to their higher prior probabilities. The similar attribute labels between different identities will influence the discriminability of the feature model for video re-id. To address this problem we apply Weighted Binary Cross Entropy (WBCE) loss:

$$\mathcal{L}_{WBCE} = -\frac{1}{L} \sum_{i=1}^L e^{\frac{1-w_i}{\sigma^2}} y_i \log(p_i) + e^{\frac{w_i}{\sigma^2}} (1-y_i) \log(1-p_i) \quad (8)$$

where  $w_i$  is the positive ratio of the  $i$ -th attribute in the training set, indicating its relative frequency. It encourages model to output rare attributes and the wrong prediction of common attributes will result in higher loss.

The attribute transfer model is trained by jointly optimizing  $\mathcal{L}_{WBCE}$  and  $\mathcal{L}_{MMD}$ . After training, the model is utilized to predict attribute labels for the re-id dataset. Specifically, for each identity, the prediction for the  $i$ -th attribute of person  $x$  is calculated by sequence merging:

$$a_i(x) = \frac{1}{T} \sum_{t=1}^T p_i(x_t) \quad (9)$$

where  $x_t$  is the  $t$ -th frame of this person, and  $p_i$  is the prediction for the  $i$ -th attribute. The  $i$ -th voted attribute label of person  $x$  is obtained by binarization:

$$L_i(x) = \begin{cases} 1 & a_i(x) \geq th \\ 0 & a_i(x) < th \end{cases} \quad (10)$$

where  $th$  is the threshold of binarization and we set  $th = 0.5$  in our method. The transferred labels are utilized as the ground truth for feature disentangling and temporal aggregation as aforementioned.

## 4. Experiments

We evaluate our proposed model on three video-based person re-id datasets: iLIDS-VID [36], PRID2011 [14] and MARS [45]. We will first introduce the datasets and evaluation metric, and then present the effectiveness of each components of our method. After comparisons with state-of-the-art methods, some qualitative results will be presented.

## 4.1. Experiment Settings

**Datasets.** The **iLIDS-VID** dataset consists of 600 image sequences of 300 identities appearing in 2 cameras. The sequence length ranges from 23 to 192 frames with an average number of 73 frames. The bounding boxes are human annotated and the challenge is mainly due to occlusion. The **PRID2011** dataset contains 2 cameras with 385 identities in camera A and 749 identities in camera B. As previous works

Sequence	BCE loss	WBCE loss	WBCE + MMD loss
	Male, Age31-45 BlackHair ub-Shirt, ub-Black lb-LongTrousers low-Black	Male, Age31-45 BlackHair ub-Jacket, ub-Black, up-Gray lb-LongTrousers low-Black shoes-White	Male, Age31-45 BlackHair ub-Shirt, ub-Black, ub-Gray lb-LongTrousers low-Black shoes-White, shoes-Sports attach-Backpack
	Male, Age31-45 BlackHair ub-Shirt, ub-White lb-LongTrousers	Male, Age17-30 BlackHair ub-Shirt, ub-White low-Gray lb-LongTrousers shoes-Black attach-Other	Male, Age17-30 BlackHair ub-Shirt, ub-White low-Gray shoes-Black attach-Backpack

Figure 5. Attribute transfer results of models trained by different loss.

we use the 200 identities appear in both cameras. The length of sequence varies from 5 to 675. The bounding boxes are also annotated by human. The **MARS** dataset is a newly released large scale dataset consisting of 1261 identities and 20715 track-lets under 6 cameras. The bounding boxes are produced by DPM detector [9] and GMMCP tracker [7]. Many sequences are of poor quality due to the failure of detection or tracking, increasing the difficulty of this dataset, which is close to real-world applications.

The attribute transfer model is trained on **RAP** [16], a large-scale pedestrian attribute dataset which provides 91 fine-grained binary attributes for each image. We choose 68 id-specific attributes (*e.g.* *BlackHair*, *TShirt*) and discard other image-specific attributes (*e.g.* *Talking*, *faceRight*). The 68 attributes are divided into 6 groups as in Table. 1.

**Evaluation metrics.** The standard protocols are performed for evaluation on these three datasets. For iLIDS-VID and PRID2011 dataset, we randomly split the dataset half-half for training and testing. The experiments are repeated 10 times with different splits and the results are averaged for stable evaluation. For MARS dataset, we follow the predefined train/test split by the original authors. 625 identities are used for training and the remaining for testing. We use the Cumulative Matching Characteristic (CMC) curve and Mean Average Precision (mAP) to evaluate the performance. The CMC value represents the average true matching being found within the first  $n$  query results. The average precision (AP) for each query is computed from its precision-recall curve. The mAP is calculated as the mean value of average precisions across all queries.

**Experiment setting.** For the network architecture, we choose ResNet-18 [12] pre-trained on ImageNet ILSVRC-2012 [28]. Input images are first resized to  $144 \times 288$  and cropped at  $128 \times 256$ . For the data augmentation, we use random crops with random horizontal mirroring for training and a single center crop for testing. We use SGD to train our model and the batch size is 32. The learning rate starts from 0.05 and is divided by 10 every 40 epochs to train the model for 100 epochs. The sequence length is set to  $T = 8$ .

Table 2. Person re-id results with different attribute transfer models. The Rank-1 CMC accuracies and mAP scores are reported.

Dataset loss	MARS		iLIDS	PRID
	mAP	R-1	R-1	R-1
BCE	67.4	81.0	78.7	89.7
BCE+MMD	70.0	81.7	79.9	90.6
WBCE	69.2	81.5	80.3	90.3
WBCE+MMD	<b>71.2</b>	<b>82.6</b>	<b>81.5</b>	<b>91.7</b>

## 4.2. Ablation Studies

**Attribute Transfer.** As aforementioned, the attribute transfer models are trained by jointly optimizing Weighted BCE loss  $\mathcal{L}_{WBCE}$  and MMD loss  $\mathcal{L}_{MMD}$ . Fig. 5 displays two examples of attribute transfer results from RAP to MARS dataset. Due to the unbalanced label distribution, the model trained by  $\mathcal{L}_{BCE}$  prefers to output common attributes, which provide little discriminative information for identification. With the variant weights corresponding to positive ratio, the  $\mathcal{L}_{WBCE}$  model is encouraged to predict unusual attributes. This model enriches the diversity of prediction results and produces important local attributes (*e.g.* shoes and attachments). However, the attributes predicted by the model trained with  $\mathcal{L}_{WBCE}$  only are not exact due to the non-ignorable domain gap between the attribute dataset and re-id dataset. Hence we propose  $\mathcal{L}_{MMD}$  to regularize the feature distribution and filter out some noise attributes.

It is hard to evaluate the attribute recognition accuracy on the re-id dataset without ground-truth labels, while the advantage of  $\mathcal{L}_{WBCE}$  and  $\mathcal{L}_{MMD}$  can be indirectly proven by the quantitative re-id accuracy, as shown in Table. 2. The person re-identification models trained with attributes transferred by  $\mathcal{L}_{WBCE}$  outperforms ones that trained by  $\mathcal{L}_{BCE}$ , both with or without  $\mathcal{L}_{MMD}$ . We attribute the improvements to the discriminative attributes produced by  $\mathcal{L}_{WBCE}$ . The joint training with  $\mathcal{L}_{MMD}$  provides consistent boost, and improves the mAP score on MARS dataset by 2.6% and 2.0% with  $\mathcal{L}_{BCE}$  and  $\mathcal{L}_{WBCE}$  respectively. The improvements demonstrate the superiority of regularizing the feature distributions between attribute dataset and re-id dataset.

**Feature disentangling and temporal aggregation.** Temporal re-weighting on disentangle features will be discussed in this section. Comprehensive experiments are performed and the results are displayed in Table. 4. Model A is the baseline model which learns feature embedding only with softmax loss and the features from different frames are merged by average pooling. Avg-pooling is a common practice for temporal aggregation in video-based re-id methods and shows competitive results [4, 10, 22]. Using map-pooling will led to about 10% point decrease in mAP on MARS. L2-normalization is also important for softmax-based method and have been chosen as an effi-

Table 3. Comparisons of our proposed approach to the state-of-the-art methods. “-” means customized networks. RGB-Only(**RO**) means that the method requires RGB frames only without optical flow for input. **SP** represents the method extract features by Single-Pass, instead of pairwise comparison for verification.

Method	settings			MARS				iLIDS			PRID		
	backbone	<b>RO</b>	<b>SP</b>	mAP	R-1	R-5	R-20	R-1	R-5	R-20	R-1	R-5	R-20
ASTPN [40]	-			-	44.0	70.0	81.0	62.0	86.0	98.0	77.0	95.0	99.0
Joint-ST [40]	CaffeNet	✓		50.7	70.6	90.0	97.6	55.2	86.5	97.0	79.4	94.4	99.3
Seq-Decision [41]	Inception	✓		-	71.2	85.7	94.3	60.2	84.7	95.2	85.2	97.1	99.6
Set2set [22]	GoogleNet	✓	✓	51.7	73.7	84.6	91.6	68.0	86.8	97.4	90.3	98.2	<b>100.0</b>
k-reciprocal [46]	CaffeNet	✓		58.0	67.8	-	-	-	-	-	-	-	-
Ours	Res-18	✓	✓	<b>71.2</b>	<b>82.6</b>	<b>93.2</b>	<b>97.7</b>	<b>82.0</b>	<b>94.3</b>	<b>98.5</b>	<b>91.7</b>	<b>98.8</b>	<b>100.0</b>
k-reciprocal [46]	Res-50	✓		68.5	73.9	-	-	-	-	-	-	-	-
TriNet [13]	Res-50	✓	✓	67.7	79.8	91.4	-	-	-	-	-	-	-
DRSA [17]	Res-50	✓	✓	65.8	82.3	-	-	80.2	-	-	93.2	-	-
Snippet [4]	Res-50			76.1	86.3	94.7	98.2	85.4	96.7	99.5	93.0	99.3	<b>100.0</b>
Ours	Res-50	✓	✓	<b>78.2</b>	<b>87.0</b>	<b>95.4</b>	<b>98.7</b>	<b>86.3</b>	<b>97.4</b>	<b>99.7</b>	<b>93.9</b>	<b>99.5</b>	<b>100.0</b>

Table 4. Person re-id results with different model settings. The Rank-1 CMC accuracies and mAP scores are reported. **F** denotes Feature disentangling. **A** means Attribute recognition. **T** represents Temporal aggregation with attribute confidence.  $\sigma$  is the hyper-parameter controlling the degree of re-weighting in Eq. 2

	settings				MARS		iLIDS		PRID	
	<b>F</b>	<b>A</b>	<b>T</b>	$\sigma$	mAP	R-1	R-1	R-1	R-1	R-1
A				-	66.1	79.0	78.3	86.8		
B	✓			-	68.2	80.1	79.9	88.6		
C	✓	✓		0.5	69.5	81.1	80.9	90.2		
D	✓	✓		-	70.3	81.7	80.4	90.9		
E	✓	✓	✓	0.3	70.6	<b>82.6</b>	81.4	91.5		
F	✓	✓	✓	0.5	<b>71.2</b>	<b>82.6</b>	<b>82.0</b>	<b>91.7</b>		
G	✓	✓	✓	0.9	71.0	82.1	81.3	91.3		
H	✓	✓	✓	1.2	70.5	81.9	80.9	90.9		

cient baseline for image-based re-id [47, 42]. Model B slightly outperforms model A with an additional attributes recognition loss. This improvement has been shown in previous works of multi-task learning. Based on model B, model C calculates the *feature level* temporal weights by the attribute recognition confidence. The improvements (1.0%/1.0%/1.6% on Rank-1 CMC) demonstrate that all frames in sequence are not equally informative and the attribute confidence provides effective evidence for temporal quality. The comparison between model B and model D shows that semantic disentangling increases the discriminability of feature representation. Model F combines model D with *sub-feature level* temporal aggregation and achieves the best results, outperforming baseline by 5.1% on mAP and 3.6%/3.7%/4.9% on Rank-1 CMC. The sub-features of one frame corresponding to different semantics ought not to share same weights due to pose changing and

occlusions, hence we further refine the temporal weights into sub-feature level. The improvement is also evident by the comparison between model F and model C.

We also carry out experiments to investigate the effect of the hyper-parameter  $\sigma$  of Eq. 2.  $\sigma$  controls the degree of temporal re-weighting. The larger the  $\sigma$  is, the smaller the variance of the temporal weights will be. The results of model E/F/G/H shows that smaller  $\sigma$  (e.g. 0.3 or 0.5) performs better by means of the high variance of temporal weights. Larger  $\sigma$  (model H) results in almost equal weights and the performance is close to average fusion (model D). We fix  $\sigma = 0.5$  and choose model F as the final result of our proposed method.

### 4.3. Comparison with the State-of-the-art Methods

In this section, the proposed approach is compared with state-of-the-art methods, and the results are displayed in Table. 3. ASTPN [40] designed a joint spatial and temporal attention pooling network. Joint-ST [40] proposed a joint spatial and temporal RNN model for video re-id. Seq-Decision [41] introduced a reinforcement learning method for pairwise decision making. Set2set [22] proposed a quality-aware network for sequence recognition. K-reciprocal [46] utilized feature encoding to address the re-ranking problem. TriNet [13] elaborated the triplet loss to train the feature model. DRSA [17] is a spatiotemporal attention model with diversity regularized item. Snippet [4] proposed competitive similarity aggregation and co-attentive snippet embedding for video re-id.

The backbone models of each method are listed for comprehensive comparison. ASTPN designed customized networks, the capability of which is comparable with Inception, CaffeNet and ResNet-18. We also display the results of our method with ResNet-50 for fair comparison with other methods. On both two levels of network capability, our

images	1	2	3	4	5	6	7	8
gender&age	0.1096	0.1252	0.1397	0.1260	0.1378	0.1373	0.1185	0.1059
head-shoulder	0.1281	0.1300	0.1248	0.1239	0.1342	0.1345	0.1123	0.1122
up-body	0.1313	0.1250	0.1313	0.1366	0.1279	0.1323	0.1078	0.1079
low-body	0.1175	0.1648	0.1553	0.1182	0.1136	0.1288	0.1084	0.1034
shoes	0.1061	0.1517	0.1554	0.1290	0.1129	0.1300	0.1055	0.1094
attach	0.1198	0.1213	0.1424	0.1234	0.1351	0.1260	0.1186	0.1134

images	1	2	3	4	5	6	7	8
gender&age	0.1438	0.1120	0.1043	0.1343	0.1204	0.1439	0.1293	0.1121
head-shoulder	0.1223	0.1337	0.1187	0.1339	0.1168	0.1214	0.1304	0.1228
up-body	0.1186	0.1227	0.1188	0.1573	0.1255	0.1165	0.1185	0.1270
low-body	0.1399	0.1352	0.0979	0.1276	0.1349	0.1156	0.1345	0.1143
shoes	0.1214	0.1377	0.1148	0.1345	0.1230	0.1222	0.1166	0.1300
attach	0.1152	0.1090	0.1166	0.1303	0.1309	0.1181	0.1419	0.1376

images	1	2	3	4	5	6	7	8
gender&age	0.1260	0.1244	0.1257	0.1256	0.1297	0.1218	0.1220	0.1248
head-shoulder	0.1244	0.1292	0.1265	0.1198	0.1256	0.1217	0.1282	0.1246
up-body	0.1292	0.1227	0.1292	0.1190	0.1236	0.1236	0.1284	0.1244
low-body	0.1326	0.1256	0.1251	0.1179	0.1263	0.1187	0.1319	0.1219
shoes	0.1273	0.1311	0.1115	0.1166	0.1255	0.1314	0.1337	0.1231
attach	0.1358	0.1286	0.1192	0.1321	0.1288	0.1160	0.1136	0.1248

Figure 6. Temporal weights of sub-features from 3 sequence. Red represents large weight and white means small weight. Best viewed in color.

methods attains the highest performance for each datasets. The superiority is prominent on the small level backbone, which improves the mAP and CMC-1 by 13.2% and 14.8% respectively on MARS dataset. Our method also boost the mAP on MARS by 2.1% with the Res-50 backbone.

In real-world applications, computational efficiency is equally important to performance. It is worth noting that Snippet and ASTPN require optical flow as input to provide motion features. However, the calculation of optical flow is very time-consuming and is hard to be applied in real-time system. Some existing methods perform pairwise comparison to calculate the similarity between query and gallery sequences, *e.g.* a pair of sequence are input to the network for verification. This strategy is impractical in large-scale scenarios because all the gallery sequences need to be calculated once for each query. An efficient practice is extracting features of large gallery set once in an off-line way and sorting them by Euclidean distances in feature space when given a query sequence. Our proposed method, which does not require optical flow and pairwise comparison, is more suitable for real-world applications. Based on the same “Res50 + RGB-Only + Sing-Pass” setting, our method significantly improves the mAP on MARS by 10.5% and boosts the CMC-1 by 4.7%/6.1%/0.7% on the three dataset.

#### 4.4. Temporal Weights Visualization

In order to demonstrate our proposed method more intuitively, the temporal weights of three sequence from MARS dataset is visualized in Fig. 6. In the first sequence, the weights of frame-7 and frame-8 are relatively small due to the occlusion and poor detection. The weights of *low-body* and *shoes* group focus in frame-2 and frame-3 due to the occlusions in other frames. Frame-3 has the highest weight in *attach* group because of the obvious bag in this frame. In the second sequence, the weights of frame-3 are small due to the occlusion, as well as frame-6. While frame-6 has the highest weight in *gender&age* group because the frontal face is visible in this frame. Frame-4 has the highest weight in *up-body* group because of the standard pose and few self-occlusion. The variance of weights in *shoes* group is small because the shoes attributes are hard to predict in all frames. The sub-feature weights in the third sequence are relatively uniform because the viewpoint and pose do not dramatically change through the sequence.

The results demonstrate that the attribute prediction confidence reflects the information quantity of each frame. The temporal weights ought to be further split into sub-feature level, because different parts are not equally informative in the sequence due to the occlusion, viewpoints and pose changing.

## 5. Conclusion

In this paper, we propose a novel algorithm of feature disentangling and temporal aggregation for video-based person re-identification. An attribute-driven method is proposed for feature disentangling, based on which we further develop sub-feature re-weighting with attribute recognition confidence, maximizing the informative regions of different frames. Moreover, a transfer learning method is introduced for automatically annotating attribute labels. Extensive experimental results on three datasets demonstrate the advantage of the proposed model over the compared state-of-the-art methods.

## 6. Acknowledgements

This paper is partially supported by NSFC (No. 61772330, 61533012, 61876109), the Basic Research Project of Innovation Action Plan (16JC1402800), the advanced research project (no.61403120201), Shanghai authentication key Lab. (2017XCWZK01), and Technology Committee the interdisciplinary Program of Shanghai Jiao Tong University (YG2015MS43).

## References

- [1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *CVPR*, pages 3908–3916, 2015.

[2] Thomas Berg and Peter Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, pages 955–962, 2013.

[3] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV*, pages 1543–1550. IEEE, 2011.

[4] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*, pages 1169–1178, 2018.

[5] Shi-Zhe Chen, Chun-Chao Guo, and Jian-Huang Lai. Deep ranking for person re-identification via joint representation learning. *IEEE TIP*, 25:2353–2367, 2016.

[6] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, pages 1335–1344, 2016.

[7] Afshin Dehghan, Shayan Modiri Assari, and Mubarak Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *CVPR*, pages 4091–4099, 2015.

[8] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015.

[9] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE T-PAMI*, 32(9):1627–1645, 2010.

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

[11] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2007.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[14] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011.

[15] Igor Kvятковский, Amit Adam, and Ehud Rivlin. Color invariants for person reidentification. *IEEE T-PAMI*, 35(7):1622–1634, 2013.

[16] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016.

[17] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, pages 369–378, 2018.

[18] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deep-reid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014.

[19] Shengcui Liao and Stan Z Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, pages 3685–3693, 2015.

[20] Yutian Lin, Liang Zheng, Zhenzhong Zheng, Yu Wu, and Yi Yang. Improving person re-identification by attribute and identity learning. *arXiv preprint arXiv:1703.07220*, 2017.

[21] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person re-identification: What features are important? In *ECCV*, pages 391–401. Springer, 2012.

[22] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *CVPR*, volume 2, page 8, 2017.

[23] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015.

[24] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015.

[25] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2208–2217. JMLR.org, 2017.

[26] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, pages 1325–1334, 2016.

[27] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton Van Den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, pages 1846–1855, 2015.

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[29] Arne Schumann and Rainer Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–28, 2017.

[30] Chen Shen, Zhongming Jin, Yiru Zhao, Zhihang Fu, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. Deep siamese network with multi-level similarity perception for person re-identification. In *ACM MM*, pages 1942–1950. ACM, 2017.

[31] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *ECCV*, pages 475–491. Springer, 2016.

[32] Arulkumar Subramaniam, Moitreyra Chatterjee, and Anurag Mittal. Deep neural networks with inexact matching for person re-identification. In *NIPS*, pages 2667–2675, 2016.

[33] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *NIPS*, pages 1988–1996, 2014.

[34] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, pages 791–808. Springer, 2016.

[35] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, pages 135–153. Springer, 2016.

[36] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703. Springer, 2014.

[37] Zhanxiong Wang, Keke He, Yanwei Fu, Rui Feng, Yu-Gang Jiang, and Xiangyang Xue. Multi-task deep neural network for joint face recognition and facial attribute prediction. In *ICMR*, pages 365–374. ACM, 2017.

[38] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, pages 1–16. Springer, 2014.

[39] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, 2018.

[40] 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 *ICCV*, pages 4743–4752. IEEE, 2017.

[41] Jianfu Zhang, Naiyan Wang, and Liqing Zhang. Multi-shot pedestrian re-identification via sequential decision making. In *CVPR*, 2018.

[42] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, pages 1077–1085, 2017.

[43] Liming Zhao, Xi Li, Jingdong Wang, and Yueling Zhuang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, volume 8, 2017.

[44] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identification. In *CVPR*, pages 144–151, 2014.

[45] 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 *ECCV*, pages 868–884. Springer, 2016.

[46] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3652–3661. IEEE, 2017.

[47] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018.

[48] Sanping Zhou, Jinjun Wang, Jiayun Wang, Yihong Gong, and Nanning Zheng. Point to set similarity based deep feature learning for person reidentification. In *CVPR*, volume 6, 2017.

[49] 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 *CVPR*, pages 6776–6785. IEEE, 2017.