Re-identifying People in Video via Learned Temporal Attention and Multi-modal Foundation Models

Cole Hill[1,2], Florence Yellin[1], Krishna Regmi[1], Dawei Du[1], and Scott McCloskey[1]
[1]Kitware Inc., USA {firstname.lastname}@kitware.com
[2]University of South Florida, USA coleh@usf.edu

Abstract

Biometric recognition from security camera video is a challenging problem when the individuals change clothes or when they are partly occluded. Others have recently demonstrated that CLIP's visual encoder performs well in this domain, but existing methods fail to make use of the model's text encoder or temporal information available in video. In this paper, we present VCLIP, a method for person identification in videos captured in challenging poses and with changes to a person's clothing. Harnessing the power of pre-trained vision-language models, we jointly train a temporal fusion network while fine-tuning the visual encoder. To leverage the cross-modal embedding space, we use learned biometric pedestrian attribute features to further enhance our model's person re-identification (Re-ID) ability. We demonstrate significant performance improvements via experiments with the MEVID and CCVID datasets, particularly in the more challenging clothes-changing conditions. In support of this and future methods that use textual attributes for Re-ID with multimodal models, we release a dataset of annotated pedestrian attributes for the popular MEVID dataset [4].

1. Introduction

Biometric recognition has become ubiquitous in daily life, with embedded facial, fingerprint, and iris sensing in mobile devices and other electronics. However, challenges persist in recognizing individuals across temporal variations (e.g., changes in clothing) and poses with occluded faces. These limitations have spurred the development of whole-body biometric recognition algorithms that work on imagery from traditional security cameras. Whole body biometric algorithms are similar to Re-ID approaches, particularly as Re-ID is increasingly evaluated on datasets where subjects change clothing. Even when subjects don't change clothing, Re-ID is challenging due to changes in lighting, pose, occlusion, and varying scale.

Large multi-modal pre-trained models continue to demonstrate strong performance on a range of tasks [19, 21, 30, 31]. The Contrastive Language-Image Pre-training (CLIP) model [17], specifically, has recently been shown to be useful for visual Re-ID [12] even when operating exclusively with still image data. We extend the approach of CLIP-based Re-ID to video recognition with a novel attention-based fusion method, and demonstrate that adding textual attributes further improves performance by fully leveraging the multi-modal CLIP model. We evaluate the efficacy of our approach with state-of-the-art (SOTA) performance on recent video Re-ID datasets including CCVID [7] and MEVID [4] which embody challenges in pose, lighting, scale variation and clothing change. Our contributions are summarized as follows:

1. We extend CLIP-based Re-ID to video Re-ID through an attention based temporal fusion method.
2. We incorporate attribute learning into our video Re-ID method and demonstrate how textual attributes *inferred* at test time can be used to more fully leverage

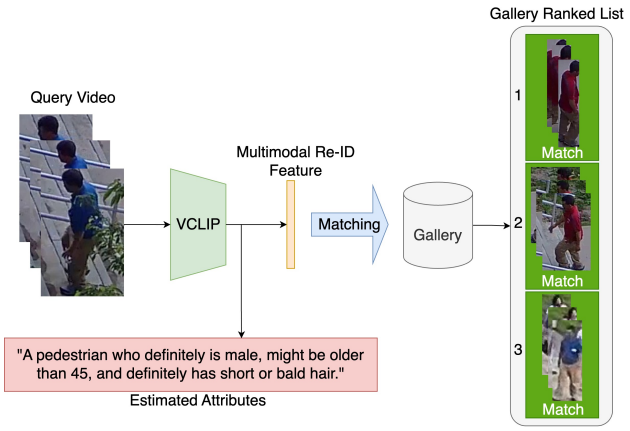

Figure 1. At inference time, our VCLIP method takes as input cropped frames from a video of an individual. It generates a Re-ID embedding based on both the person's visual appearance and textual attributes inferred from the query video. The resulting multi-modal Re-ID feature can be used to match against a gallery of videos to produce accurate ranked lists in challenging cases such as clothing changes.

CLIP's text embedding features for Re-ID. If available, our method can also utilize ground-truth attributes during training to tune the attribute prompts.

3. We introduce MEVID Person Attributes, a dataset of annotated pedestrian attributes for the challenging Re-ID dataset MEVID. Whereas the Re-ID performance on some datasets has saturated, attributes associated with them are not needed to drive improved recognition performance. We demonstrate that the attributes we provide for the more challenging MEVID dataset offer a means to achieve higher Re-ID performance.

We demonstrate the impact of these contributions with new SOTA results on two challenging clothing-changing video person identification datasets, MEVID and CCVID. We improve the mAP and CMC scores on the MEVID dataset, improving rank-1 by 1.3% and rank-20 matching by 1.9% using a few biometrically-relevant pedestrian attributes. Through ablation, we demonstrate performance improvements attributable to these contributions. We also demonstrate, via an experiment with ground truth attributes, that further improvements to pedestrian attribute inference could contribute to better mAP and CMC scores.

2. Related Work

Video Person Re-Identification. Video-based person Re-ID methods [9, 10, 25, 26, 29] rely on spatial-temporal appearance representation learning. To deal with cloth-changing situations, Gu *et al*. [7] design a Clothes-based Adversarial Loss to mine clothes-irrelevant features by penalizing the predictive power of the Re-ID model w.r.t. clothes. Since appearance information is not reliable for cloth-changing pedestrians, other biometric cues are also considered for training networks, such as silhouette, 3-D body shape and skeletons sequences. For example, Nguyen *et al*. [15] first captured the temporal dynamics from video sequences and then estimated frame-wise shape parameters by an identity-aware 3D regressor. Zhu *et al*. [33] developed a Pose and Shape Encoder to model body shape and an Aggregated Appearance Encoder to fuse temporal appearance features. In contrast, we use biometric pedestrian attribute features to distinguish between different subjects with changing clothes.

Vision-language learning. Recently, researchers leverage vision-language learning (*i.e*., CLIP [17, 32]) to extract more generic appearance representations [3, 12, 28]. By first learning a set of discriminative text tokens for each ID, CLIP-ReID [12] fine-tunes the image encoder of CLIP [17] to extract Re-ID embeddings constrained within the rich CLIP feature space. Chen *et al*. [3] apply a CLIP-like framework with new memory-swapping contrastive learning to visible-infrared person Re-ID. Different from the above works focusing on image Re-ID, TF-CLIP [28] extends the CLIP model for video person Re-ID by proposing the Temporal Memory Diffusion (TMD) module. In this work, we learn multi-level spatial-temporal features based on pre-trained language-image models to produce discriminative cross-modal embeddings.

3. Method

Given a probe video of a subject, biometric video Re-ID aims to locate that individual from a gallery of enrollments videos. Our VCLIP method, shown in Fig. 2 uses cross-modal visual, text, and attribute features to generate discriminative embeddings useful for video Re-ID in the challenging clothing-changing setting. We utilize attribute learning to fully take advantage of our model's cross-modal embedding space, and we use a co-attention mechanism to fuse attribute and visual features. Temporal fusion then combines per-frame features into per-video features. Using a multi-stage training strategy, we finetune/train all model components, resulting in the final VCLIP model used for inference.

3.1. CLIP Overview

The key to successful biometric video Re-ID is extracting biometric feature representations that are invariant to camera, domain, and clothing. We therefore propose a method for extracting cross modal spatial-temporal features from videos that capitalizes on the recent success using CLIP's [17] large pretrained vision and language models for their high quality feature representations and zero-shot transfer ability. CLIP consists of text and image encoders that compute embeddings from a pair of visual inputs and corresponding text descriptions. The text encoder $\mathcal{T}$ is a transformer network, and the visual encoder $\mathcal{V}$ is implemented as either a CNN, such as ResNet-50, or a transformer, such as ViT-B/16, which we use for our method. Given an input image x of subject with identity y, the visual encoder generates an image feature,

$$f_{\text{vis}}(x) = \mathcal{V}(x). \tag{1}$$

We also generate a per-subject identity sentence associated with each video, "A photo of y." Each identity sentence is tokenized via byte pair encoding [18], embedded into a 512-dimensional vector for each token, and encoded with the CLIP text encoder into a text feature,

$$f_{\text{text}}(y) = \mathcal{T}(y). \tag{2}$$

The model is trained to minimize the dot product distance $s(\cdot,\cdot)$ between the [EOS] token embedding of the text feature, $\hat{f}_{\text{text}}(y)$, and the [CLS] token embedding of the image feature, $\hat{f}_{\text{vis}}(x)$, both of which are first projected into a cross-modal embedding space via linear layers W_{text} and W_{vis}[1]. To address ambiguous text descriptions, CoOp [32]

[1]For linear projections *i.e*. W^T, superscript T denotes transpose.

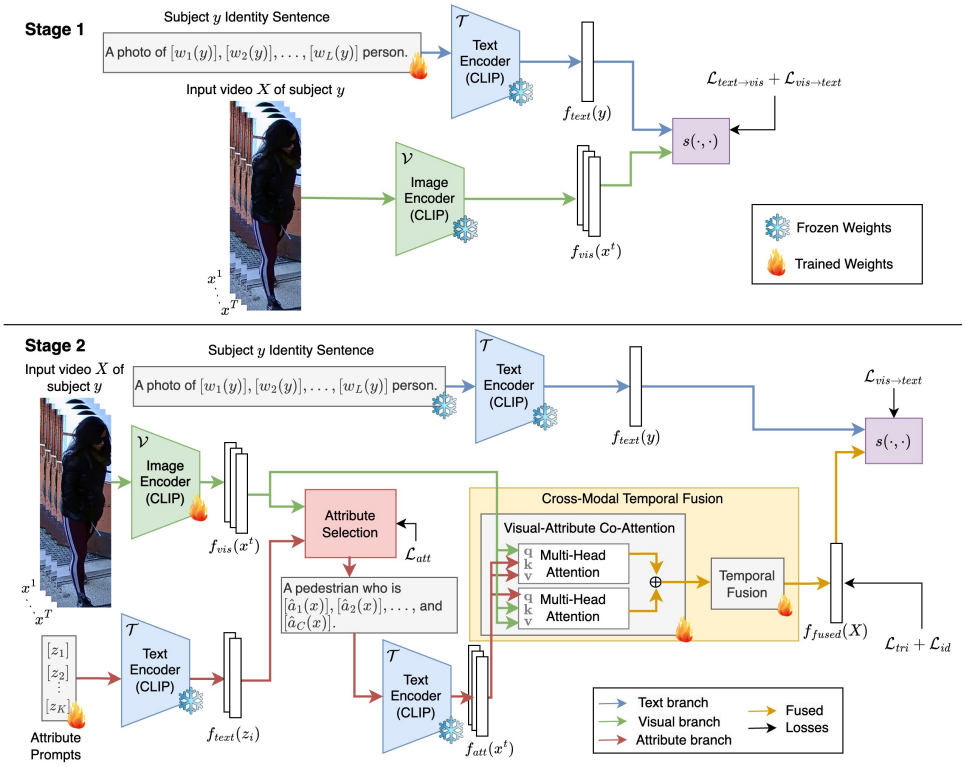

Figure 2. Overview of our approach: VCLIP utilizes learned person attribute sentences encoded with the text encoder and visual features extracted using the image encoder. The cross-modal features– used during inference– are fused using co-attention transformers. During the first stage of training, we learn subject identity prompts. During the second, we finetune the image encoder, attribute prompts (if ground truth attributes are available), and train the temporal components of our model. The text encoder remains frozen throughout all stages.

introduces learned prompt tokens, which are trained while keeping the text and image encoders fixed. CLIP-ReID [12] similarly trains learnable prompts for the Re-ID task during an initial training stage in which the text and image encoders are fixed, allowing the ambiguous subject identities to be mapped to learned prompts whose cross-modal embeddings are aligned with the subjects' visual embeddings. As in [12], we replace each identity sentence with L learned prompts, $\{w_l(y)\}_{l=1}^{L}$, to generate per-subject sentences, *i.e.*, "A photo of $[w_1(y)]$ $[w_2(y)]$... $[w_L(y)]$ person."

3.2. Attribute Learning with Prompts

An advantage of using a cross-modal model is our ability to utilize pedestrian attribute information that might be available in addition to video clips. In the event that pedestrian attributes are *unknown*, we use pedestrian attribute recognition to *infer* attributes from an input video. Pedestrian attribute recognition [6, 11, 13, 22, 23, 27] aims to describe semantic information about an individual using a set of predefined attributes, including biometric information (*e.g.*, gender, age) and clothing descriptions (*e.g.*, long sleeves, skirt, carrying backpack). We use CLIP's cross-modal embedding space to embed not only learned prompts corresponding to individual subjects, but also textual attributes describing each video.

Our attribute branch, (Fig. 2, red arrows) consists of learned attribute prompts and an attribute selection module, which are used to generate a per-frame attribute sentence. For attribute learning, we use a set of K attributes $\{a_i\}_{i=1}^{K}$ where each of the attributes belongs to one of C mutually exclusive classes $\{\mathcal{C}_i\}_{i=1}^{C}$. For example, we might use $K = 9$ attributes (*i.e.* male, female, age 16 to 31, ages 31 to 45, age 45 to 60, above 61, long hair, short hair, and bald) and $C = 3$ attribute classes (*i.e.* gender, age, and hairstyle). Within each attribute class an individual must be assigned to one and only one attribute (*i.e.* an individual must belong to exactly one age group). That is to say with attributes $a_i, a_j \in \mathcal{C}_n$ an individual must have attribute a_i or a_j but not both. Using these attributes we learn K attribute prompts $\{z_i\}_{i=1}^{K}$, where each prompt corresponds to a unique attribute. Note that if ground truth attributes are not available, we keep the attribute prompts frozen.

We then compute the cosine similarity between the visual feature $f_{\text{vis}}(x)$ and the text feature for each of the attribute prompts $f_{\text{text}}(z_i)$. Each attribute is then assigned a probability based by applying the softmax function over

each attribute class

$$p_{\text{att}}(a_i(x)) = \frac{e^{s(f_{\text{vis}}(x), f_{\text{text}}(z_i))}}{\sum_{a_k \in \mathcal{C}_n} e^{s(f_{\text{vis}}(x), f_{\text{text}}(z_k))}}, \quad (3)$$

so that the probabilities of all attributes within a single class sum to one (*i.e.* $\sum_{a_i \in \text{age class}} p_{\text{att}}(a_i) = 1$).

For each class i and image frame x, we select the most probable attribute $\hat{a}_i(x)$ to generate an attribute sentence, *i.e.*, "A pedestrian who is $[\hat{a}_1(x)]$, $[\hat{a}_2(x)]$, ..., and $[\hat{a}_C(x)]$." Because our goal is person Re-ID in the more challenging *clothing changing* setting [4, 7], we eliminate clothing or action related attributes and retain only *biometric* attributes. When using predicted attributes we also experiment with adding confidence to the attribute sentence by adding phrases which imply doubt when the attribute probability $p_{att}(a_i(x))$ is low. Specifically, when the attribute probability is less than 75% we use the word "might" and when greater than 75% we use the word "definitely", producing sentences such as "A pedestrian who might be $[\hat{a}_1(x)]$, might be $[\hat{a}_2(x)]$, ..., and definitely is $[\hat{a}_C(x)]$]."

Each attribute sentence is of length 77 and is tokenized with the byte pair encoding (BPE) [18], embedded into a 512-dimensional vector for each token, and encoded into a feature in the cross-modal embedding space with the CLIP text encoder. The final feature produced by the attribute branch of our model for input frame x is denoted $f_{att}(x)$.

3.3. Cross-Modal Temporal Fusion

In addition to incorporating attribute learning for Re-ID, our method extends CLIP to use *video* inputs, instead of single *frame* inputs. Given an input video X with T frames, $X = \{x^t\}_{t=1}^T$ of subject y, we generate an image feature and an attribute feature for each frame $t = 1, 2, \ldots, T$ in the video. The attribute features $\{f_{att}(x^t)\}_{t=1}^T$ are fused with the image features $\{f_{\text{vis}}(x^t)\}_{t=1}^T$ via cross-modal temporal fusion, shown in Fig. 2 (yellow arrows indicate fused features). The fusion module generates for each video a feature $f_{\text{fused}}(X)$ whose embedding aligns with the video's subject identity feature $f_{\text{text}}(y)$.

Given attribute and visual features for each frame, we use a cross-modal transformer with attention to achieve a unified feature representation. Rather than fusing the modalities with concatenation or summation fusion, as in [2], we use cross-modal attention to fuse the attribute and visual branches. Using learned attention for cross modal fusion allows us to fully take advantage of complimentary information extracted from the various modalities and to fuse them into a discriminative, task-aware feature.

We use co-attention transformers [14], originally designed to fuse vision and language features, to fuse visual features and attribute features by exchanging key-value pairs in multi-headed attention layers. To fuse the attribute and visual features, we use a co-attention transformer $CO(\mathbf{q}, \mathbf{k}, \mathbf{v})$ with the attribute features as queries $\mathbf{q}$ and the final layer of the CLIP visual encoder projected into 512-dimensions via W_{vis} as keys $\mathbf{k}$ and values $\mathbf{v}$,

$$\mathbf{q} = f_{\text{att}}(x^t), \quad \mathbf{k} = f_{\text{vis}}(x^t)W_{\text{vis}}^T, \quad \mathbf{v} = f_{\text{vis}}(x^t)W_{\text{vis}}^T, \quad (4)$$

and visa versa,

$$\mathbf{q} = f_{\text{vis}}(x^t)W_{\text{vis}}^T, \quad \mathbf{k} = f_{\text{att}}(x^t), \quad \mathbf{v} = f_{\text{att}}(x^t). \quad (5)$$

To combine features from each frame x^t into a single per-video feature, we utilize a temporal transformer. A learned positional encoding is added to the per frame fused features. These features are then fed into a two-layer transformer, and the output is mean pooled along the temporal dimension to produce the final fused video feature $f_{\text{fused}}(X)$.

3.4. Training Strategy

We employ a two stage training strategy to train the various components of our model. In order to utilize the high quality feature representations from the large pretrained CLIP models while also finetuning its image encoder for our task and dataset, we freeze the weights of the CLIP text encoder and progressively learn the new components of our model and finetune the CLIP image encoder.

During the first stage of training, as in [12], we learn only the identity prompts and keep the rest of the model's weights frozen. Given a batch with B videos $\{X_b\}_{b=1}^B$ of subjects $\{y_b\}_{b=1}^B$, we define the visual-to-text and text-to-visual contrastive losses in terms of the dot product similarity of projected features,

$$s_{i,j} = s\left(\hat{f}_{\text{text}}(y_i)W_{\text{text}}^T), (\hat{f}_{\text{vis}}(X_j)W_{\text{vis}}^T\right). \quad (6)$$

The visual-to-text constrastive loss is defined as the similarity between visual and text features, normalized by text features,

$$\mathcal{L}_{\text{vis}\to\text{text}}(b) = -\log\left(\frac{\exp(s_{b,b})}{\sum_{j=1}^B \exp(s_{j,b})}\right). \quad (7)$$

The text-to-visual contrastive loss is likewise defined as the similarity between visual and text features, normalized by images features,

$$\mathcal{L}_{\text{text}\to\text{vis}}(y_b) = \frac{-1}{\mathcal{N}(y_b)} \sum_{\{n|y_n=y_b\}} \log\left(\frac{\exp(s_{b,n})}{\sum_{j=1}^B \exp(s_{b,j})}\right), \quad (8)$$

where $\mathcal{N}(y_b)$ is the number of videos in the batch with identity y_b. Note that during the first stage of training, we do not utilize the attribute model to reduce computational complexity.

During the second stage of training, we finetune the CLIP image encoder while jointly learning attribute prompts, and the weights of the video-attribute fusion model and temporal fusion transformer. For our losses we use triplet loss, an attribute loss, and identity loss with label smoothing. The triplet loss is defined as

$$\mathcal{L}_{\text{tri}} = \max\left(d_{\text{pos}} - d_{\text{neg}} + \alpha, 0\right), \tag{9}$$

where d_{pos} and d_{neg} are the distances between positive and negative pairs of features, and α is the margin. We sum the triplet loss computed over several temporal features, including $\hat{f}_{\text{temp}}(X)$ and $\hat{f}_{\text{temp}}(X)W_{\text{vis}}^{T}$. The identity loss with label smoothing is

$$\mathcal{L}_{\text{ID}} = -\sum_{n=1}^{N} \gamma_n \log(\hat{y}_n), \tag{10}$$

where γ_n is the target distribution, N is the total number of unique identities, and $\hat{y}_n$ is the logit value for the n-th identity computed based on temporal features. We again sum the identity losses computed for the various features.

To facilitate optimizing the attribute prompts when ground truth attributes are available, we use an attribute loss which uses the binary cross entropy loss applied to the estimated attribute probabilities $\hat{p}_{\text{att}}(X)$, estimated per video, and the ground truth attribute vector q,

$$\mathcal{L}_{\text{att}} = -q \cdot \log\left(\hat{p}_{\text{att}}(X)\right) - (1-q) \cdot \log\left(1 - \hat{p}_{\text{att}}(X)\right). \tag{11}$$

When ground truth attributes are not available, the attribute prompts are frozen.

In summary, we use the following losses for the two stages of training:

$$\mathcal{L}_{\text{text}\to\text{vis}} + \mathcal{L}_{\text{vis}\to\text{text}} \qquad \text{stage 1} \tag{12}$$

$$\mathcal{L}^{\gamma}_{\text{vis}\to\text{text}} + \mathcal{L}_{\text{tri}} + \mathcal{L}_{\text{id}} + \mathcal{L}^{*}_{\text{att}} \qquad \text{stage 2} \tag{13}$$

*If ground truth attributes are available.

During the test time, we compute the fused features $f_{\text{fused}}(X)$ for the gallery tracklets and the query video tracklets. We use these fused features for finding the match for the query with the gallery.

4. Experimental and Implementation Details

To quantify the performance of our VCLIP method, we perform a series of experiments on video Re-ID datasets, described below. We also describe the person attributes that we have annotated in support of these experiments.

4.1. Experimental Details

Datasets To quantify the performance of the proposed method on the task of video-based person Re-ID, we conduct experiments using two different public video Re-ID datasets: MEVID [4] and CCVID [7]. These datasets represent a range of challenges including long-term Re-ID (CCVID), large scale variations (MEVID), indoor-outdoor matching (MEVID), and clothing changes (both). MEVID consists of 158 identities and CCVID has 226 unique identities in the dataset. MEVID is more challenging in the sense that it includes 33 viewpoints and multiple scales of images from 33 different settings.

Figure 3. Examples of MEVID attributes. Note for our experiments we only use *biometric* attribute labels shown in blue.

4.1.1 MEVID Person Attributes

We release a new dataset for person attribute recognition, the MEVID Person Attributes Dataset: we enrich the MEVID dataset with 40 binary and 1 multi-class manually-annotated attributes, shown in Tab. 1. MEVID is a large-scale video Re-ID dataset with diversified data collects. It consists of 158 unique people with 598 outfits in 8,092 tracklets; 4 different outfits per identity and collected with 33 camera views across 17 different camera locations both indoors and outdoors. For attributes annotations, the check-in photos of the identities are utilized. The person attributes are defined based on the following categories: sex, age range, body type, clothes details, accessories and carryings.

We employ three annotators to annotate the dataset. The conflicts in annotations are resolved using the majority voting of the annotators, *e.g.*, if two of the three annotators agree to an annotation whereas the third annotator disagree, we will keep the annotation of the first two annotators. If all three annotators disagree, the annotation is resolved by a fourth annotator who checks the corresponding image and the three annotations and makes a decision. Some example annotations with corresponding images are shown in Fig. 3.

Implementation Details For our main experiments, we use three simplified MEVID biometric classes (Tab. 1): gender (male/female), age (younger/older than 45), and hair

Table 1. All attributes (40 binary and 1 multi-class) included in the MEVID Person Attributes Dataset are listed by class category.

Class	# Binary Attributes	Attributes
Sex	2	male, female
Age category	4	16-30, 31-45, 45-60, above 61
Body type	3	underweight, healthy weight, overweight
Hair length	3	bald, short, long
Facial hair	1	-
Hair color	N/A	black, brown, blonde, red, grey, other color, N/A
Upper body clothing	8	t-shirt, collared shirt, dress, vest, suit, sweater, short coat, long coat, hoodie
Sleeve length	2	long sleeves, short sleeves
Lower body clothing	5	jeans, long skirt, short skirt, shorts, other pants
Accessories	6	hat, glasses, scarf, jewelry, gloves, no accessories
Carrying	6	handbag, crossbody bag, backpack, briefcase, carrying other, carrying nothing

length (long/short or bald). For the CCVID dataset which does not have ground truth attributes we utilize the attribute prompts learned from MEVID. We use visual encoder and text encoder from CLIP [17] as the backbone for our image and text feature extractor. Specifically, we use the ViT-B/16 model architecture. As discussed earlier, we employ a two-stage training strategy. The first stage learns the prompts for class ids. The second stage optimizes the network to learn joint textual (attributes) and visual features as well as temporal fusion. For training, we use Adam optimizer and learning rate scheduling. The first-stage training is conducted for 120 epochs. For the second stage we train for 15 epochs. We use a batch size of 32 and sequence length of 8. Frame sampling is done randomly during training and evenly during the test time. If the tracklets are shorter than 8 frames during the test, the frames are resampled until there are 8 frames used for the video tracklet.

5. Results

5.1. Evaluation Metrics

Following the prior works on video person Re-ID [7, 16, 33], we evaluate our proposed method and compare its performance with other Re-ID methods on different datasets using mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) metrics. The CMC metric scores the success in finding the true match within the first k ranks ($k = 1, 5, 10, 20$), and the mAP metric is based on precision and recall over all the queries.

We also evaluate performance based on the clothing of the identities. Here, we compute the top-k accuracy and mAP in the challenging clothes-changing (CC) setting.

5.2. Experimental Results Analysis

Baselines We compare our method to SOTA Re-ID methods, including [7, 33] and others, on the MEVID and CCVID datasets. MEVID results are shown in Tables 2, 4, and 5, and CCVID results are shown in Tab. 3. On the MEVID dataset, where there are ground truth attributes to learn attribute prompts, VCLIP performs better than the SOTA methods (Tab. 2): we see performance improvements of 1.3% in Rank-1 and 1.9% in Rank-20. On the CCVID dataset, where ground truth attributes are not available, our method is competitive with most prior work (Tab. 3).

Table 2. Comparative Analysis of proposed method over SoTA methods on MEVID Dataset [4]. The best scores are shown in **bold** whereas the second best scores are underlined.

Method	**mAP**	**Rank**			
		1	**5**	**10**	**20**
BiCnet-TKS [9]	6.3	19.0	35.1	40.5	52.9
PiT [29]	13.6	34.2	55.4	63.3	70.6
STMN [5]	11.3	31.0	54.4	65.5	72.5
AP3D [8]	15.9	39.6	56.0	63.3	76.3
TCLNet [10]	23.0	48.1	60.1	69.0	76.3
PSTA [24]	21.2	46.2	60.8	69.6	77.8
AGRL [26]	19.1	48.4	62.7	70.6	77.9
Attn-CL [16]	18.6	42.1	56.0	63.6	73.1
Attn-CL+rerank [16]	25.9	46.5	59.8	64.6	71.8
CAL [7]	27.1	52.5	66.5	73.7	80.7
ShARc [33]	**29.6**	59.5	70.3	77.2	82.9
VCLIP (Ours)	27.7	**60.8**	**73.1**	**78.2**	**84.8**

Qualitative Results Example video retrievals are shown in Fig. 4 for our method and another SOTA method CAL [7]. The top row showcases a query video from a challenging high-pitch viewpoint, where VCLIP successfully identifies tracks containing the query subject despite face occlusion, headwear, and clothing changes. In contrast, CAL appears to rely on superficial color cues from the subject's shirt, failing to accurately identify all but one track. Furthermore, the second row demonstrates VCLIP's

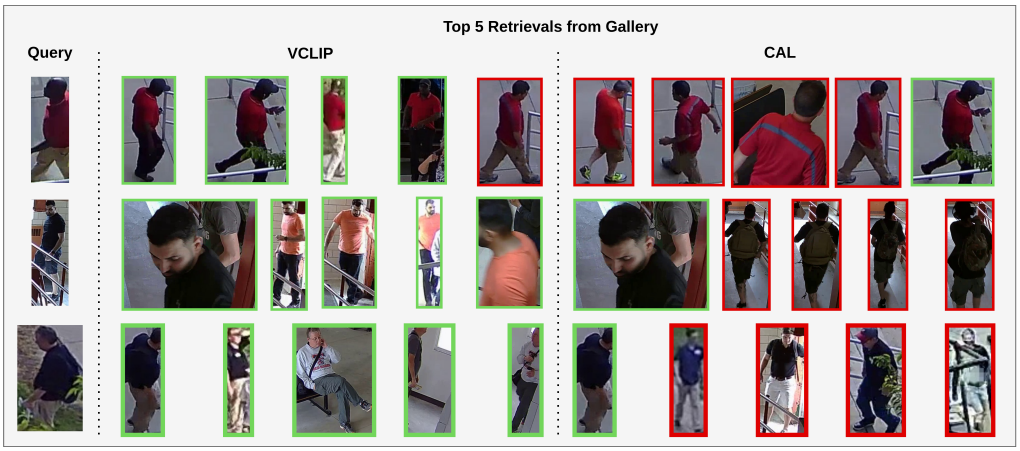

Figure 4. Qualitative samples on MEVID Dataset [4]. The first image represents a query frame, followed by the top-5 retrieved tracklets. Frames with green borders are successful retrievals and failures are shown by frames with red borders.

Table 3. Performance of VCLIP and SoTA methods on the CCVID Dataset [7]. CC refers to clothes changing setup of evaluation. The best scores are shown in **bold** whereas the second best scores are underlined.

Method	General		CC	
	mAP	Rank-1	mAP	Rank-1
GaitNet [20]	56.5	62.6	49.0	57.7
GaitSet [1]	73.2	81.9	62.1	71.0
CAL [7]	81.3	82.6	79.6	81.7
ShARc [33]	**90.2**	**89.8**	**85.2**	**84.7**
VCLIP (Ours)	86.0	86.8	81.8	84.3

ability to robustly identify subjects across different outfits, whereas CAL is again observed to be biased towards individuals wearing a similar outfit to the query video. Finally, the third row showcases VCLIP's ability to match videos across scales and between indoor and outdoor environments.

MEVID Challenges The MEVID dataset [4] was released with three challenges to facilitate an in-depth evaluation of existing Re-ID methods. We compare the performance of VCLIP (without ground truth attributes) against several SOTA Re-ID methods in the Change-of-Clothing Challenge (Tab. 4). Notably, our results demonstrate a substantial improvement over prior work, with significant boosts in rank retrieval accuracy observed at Rank-5 (+6.7%), Rank-10 (+11.3%), and Rank-20 (+20.7%). Moreover, in the Scale Variation Challenge (Tab. 5), we present our method's performance under uniform scale conditions and varying scale settings, achieving optimal or second-best mAP and Rank metrics across all scenarios. Lastly, for the Location Difference Challenge, VCLIP exhibits parity with current SOTA methods, outperforming them under same location conditions while securing a second-place finish when gallery video locations are different than the query videos. Further details can be found in the Supplementary Materials. We would like to note that the performance of ShARc [33] on these specific challenges could not be evaluated, as relevant results were not reported in their publication and source code for their method was not made publicly available.

Table 6. Ablation results for various modules in the proposed VCLIP method on the MEVID Dataset. The ablations shown are Temporal Fusion (TF), Estimated Attributes (ES), Attribute Certainty (AC), and Ground Truth Attributes (GT) [4].

TF	ES	AC	GT	mAP	Rank			
					1	5	10	20
✗	✗	✗	✗	19.8	52.2	72.2	78.2	83.2
✓	✗	✗	✗	20.0	52.5	69.3	76.6	82.0
✓	✓	✗	✗	27.4	59.5	72.5	77.8	85.8
✓	✓	✓	✗	27.7	60.8	73.1	78.2	84.8
✗	✗	✗	✓	28.0	61.7	72.2	76.3	86.7
✓	✗	✗	✓	**45.1**	**70.3**	**83.2**	**87.3**	**90.8**

Ablation Study We conducted an ablation study to evaluate the contributions of various components in our model (Tab. 6). Our experiments show that incorporating ground truth attributes during training and testing significantly improves video-based person re-identification (Re-ID) performance. Specifically, using ground truth attributes outperforms our baseline approach without attributes. Temporal fusion provides a gain in recognition performance, while predicted CLIP attributes improve Re-ID accuracy. Conditioning the attribute sentence with certainty words also enhances model performance. The results further suggest that improved attribute prediction could lead to significant further gains in Re-ID performance.

6. Ethical Considerations

Improving the performance of non-cooperative biometric recognition has the potential to increase public safety

Table 4. Comparison with prior work on the MEVID dataset for the Change-of-Clothing Challenge. The best scores are shown in **bold** whereas the second best scores are underlined.

Method	Same Clothes					Different Clothes				
	mAP	Rank				mAP	Rank			
		1	5	10	20		1	5	10	20
BiCnet-TKS [9]	8.0	0.5	36.5	41.7	51.4	1.7	0.7	4.6	7.8	13.4
PiT [29]	19.5	36.8	58.7	66.3	73.6	2.0	1.1	5.3	8.5	13.7
STMN [5]	18.5	33.7	58.3	69.1	76.4	1.2	0.4	1.8	3.9	6.0
AP3D [8]	23.2	42.7	59.7	67.7	79.2	2.9	1.8	7.4	9.5	16.6
TCLNet [10]	31.9	51.7	63.5	71.9	79.2	3.9	3.5	8.8	14.1	21.1
PSTA [24]	29.7	49.0	63.9	72.2	78.5	5.1	5.6	12.3	19.4	28.9
AGRL [26]	32.6	51.4	64.9	73.6	80.9	5.7	4.9	15.1	19.0	25.7
Attn-CL [16]	24.2	44.4	59.7	66.3	72.6	3.4	2.8	8.5	15.5	24.6
Attn-CL+rerank [16]	34.1	50.7	63.2	68.1	72.9	4.2	2.1	9.2	13.7	22.5
CAL [7]	**39.0**	56.6	70.8	78.1	85.4	4.3	3.5	10.6	14.8	19.4
VCLIP(Ours)	37.6	**63.5**	**76.4**	**81.9**	**87.2**	**6.7**	**6.7**	**17.3**	**26.1**	**40.1**

Table 5. Comparison with prior work on the MEVID dataset for the Scale Variation Challenge. The best scores are shown in **bold** whereas the second best scores are underlined.

Method	Same Scale					Different Scale				
	mAP	Rank				mAP	Rank			
		1	5	10	20		1	5	10	20
BiCnet-TKS [9]	5.2	14.7	26.0	31.0	39.7	4.6	10.7	20.8	25.5	34.9
PiT [29]	11.4	23.7	44.0	53.3	60.3	10.6	23.5	37.3	41.6	51.0
STMN [5]	10.5	19.2	33.2	39.2	46.2	9.4	22.3	42.0	51.0	58.0
AP3D [8]	14.2	31.0	47.3	53.0	63.7	11.4	24.5	35.9	42.6	52.7
TCLNet [10]	20.7	40.0	52.3	61.0	65.0	17.6	34.2	44.3	50.3	59.4
PSTA [24]	18.6	34.3	51.3	60.3	67.0	16.8	29.9	44.3	51.7	60.1
AGRL [26]	22.1	40.3	57.3	64.7	71.0	17.7	29.5	43.6	49.3	58.4
Attn-CL [16]	15.4	31.3	50.0	56.0	64.3	14.3	26.5	37.6	46.3	55.7
Attn-CL+rerank [16]	23.1	35.7	54.0	59.3	67.7	21.2	33.6	44.0	50.0	54.4
CAL [5] [7]	**24.3**	42.3	56.3	61.3	71.7	20.6	35.2	50.7	**58.4**	62.4
VCLIP (Ours)	24.0	**49.7**	**66.7**	**74.0**	**78.3**	**21.5**	**39.6**	**52.3**	58.1	**67.4**

but, if used by malicious actors, could lead to harassment, persecution, or other bad outcomes.We note that the MEVID dataset was collected under IRB supervision and that all people depicted in this paper expressly consented to have their likeness appear in publications.

7. Conclusion

We extend the application of CLIP-based Re-ID to video through attention-based temporal fusion methods, yielding improved performance compared to dense temporal approaches [33]. Additionally, we propose a novel extension that leverages textual attributes within the multi-modal CLIP framework, enabling more effective person identification. Our approach achieves SOTA results on the MEVID dataset with significant advances in robustness to clothing changes.

We observe a notable gap between performance using ground truth attributes and predicted attributes, indicating opportunities for further improvement in attribute prediction and the exploitation of textual attributes under weak/noisy supervision.

8. Acknowledgemet

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract 2022-21102100003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8126–8133, Jul. 2019. 7

[2] Cuiqun Chen, Mang Ye, and Ding Jiang. Towards modality-agnostic person re-identification with descriptive query. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15128–15137, 2023. 4

[3] Zhong Chen, Zhizhong Zhang, Xin Tan, Yanyun Qu, and Yuan Xie. Unveiling the power of clip in unsupervised visible-infrared person re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3667–3675, 2023. 2

[4] Daniel Davila, Dawei Du, Bryon Lewis, Christopher Funk, Joseph Van Pelt, Roderic Collins, Kellie Corona, Matt Brown, Scott McCloskey, Anthony Hoogs, and Brian Clipp. Mevid: Multi-view extended videos with identities for video person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1634–1643, January 2023. 1, 4, 5, 6, 7

[5] Chanho Eom, Geon Lee, Junghyup Lee, and Bumsub Ham. Video-based person re-identification with spatial and temporal memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12036–12045, 2021. 6, 8

[6] Xinwen Fan, Yukang Zhang, Yang Lu, and Hanzi Wang. Parformer: Transformer-based multi-task network for pedestrian attribute recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 3

[7] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *CVPR*, 2022. 1, 2, 4, 5, 6, 7, 8

[8] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 228–243, Cham, 2020. Springer International Publishing. 6, 8

[9] Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, and Shiguang Shan. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2014–2023, June 2021. 2, 6, 8

[10] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 388–405, Cham, 2020. Springer International Publishing. 2, 6, 8

[11] Jian Jia, Xiaotang Chen, and Kaiqi Huang. Spatial and semantic consistency regularizations for pedestrian attribute recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 962–971, 2021. 3

[12] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1405–1413, 2023. 1, 2, 3, 4

[13] Wanhua Li, Zhexuan Cao, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Label2label: A language modeling framework for multi-attribute learning. In *European Conference on Computer Vision*, pages 562–579. Springer, 2022. 3

[14] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 4

[15] Vuong D Nguyen, Pranav Mantini, and Shishir K Shah. Temporal 3d shape modeling for video-based cloth-changing person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 173–182, 2024. 2

[16] Priyank Pathak, Amir Erfan Eshratifar, and Michael Gormish. Video person re-id: Fantastic techniques and where to find them (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(10):13893–13894, Apr. 2020. 6, 8

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 6

[18] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015. 2, 4

[19] Shuai Shao, Yu Bai, Yan Wang, Baodi Liu, and Yicong Zhou. Deil: Direct-and-inverse clip for open-world few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28505–28514, 2024. 1

[20] Chunfeng Song, Yongzhen Huang, Yan Huang, Ning Jia, and Liang Wang. Gaitnet: An end-to-end network for gait based human identification. *Pattern Recognition*, 96:106988, 2019. 7

[21] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13171–13182, 2024. 1

[22] Chufeng Tang, Lu Sheng, Zhaoxiang Zhang, and Xiaolin Hu. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4997–5006, 2019. 3

[23] Xiao Wang, Jiandong Jin, Chenglong Li, Jin Tang, Cheng Zhang, and Wei Wang. Pedestrian attribute recognition via clip based prompt vision-language fusion. *arXiv preprint arXiv:2312.10692*, 2023. 3

[24] Yingquan Wang, Pingping Zhang, Shang Gao, Xia Geng, Hu Lu, and Dong Wang. Pyramid spatial-temporal aggregation

for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12026–12035, October 2021. 6, 8

[25] Jinlin Wu, Lingxiao He, Wu Liu, Yang Yang, Zhen Lei, Tao Mei, and Stan Z. Li. Cavit: Contextual alignment vision transformer for video object re-identification. In *ECCV*, 2022. 2

[26] Yiming Wu, Omar El Farouk Bourahla, Xi Li, Fei Wu, Qi Tian, and Xue Zhou. Adaptive graph representation learning for video person re-identification. *IEEE Transactions on Image Processing*, 29:8821–8830, 2020. 2, 6, 8

[27] Jie Yang, Jiarou Fan, Yiru Wang, Yige Wang, Weihao Gan, Lin Liu, and Wei Wu. Hierarchical feature embedding for attribute recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13055–13064, 2020. 3

[28] Chenyang Yu, Xuehu Liu, Yingquan Wang, Pingping Zhang, and Huchuan Lu. Tf-clip: Learning text-free clip for video-based person re-identification. In *AAAI*, 2023. 2

[29] Xianghao Zang, Ge Li, and Wei Gao. Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. *IEEE Transactions on Industrial Informatics*, 18(12):8776–8785, Dec. 2022. 2, 6, 8

[30] Bingfeng Zhang, Siyue Yu, Yunchao Wei, Yao Zhao, and Jimin Xiao. Frozen clip: A strong backbone for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3796–3806, 2024. 1

[31] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022. 1

[32] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2

[33] Haidong Zhu, Wanrong Zheng, Zhaoheng Zheng, and Ram Nevatia. Sharc: Shape and appearance recognition for person identification in-the-wild. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6290–6300, January 2024. 2, 6, 7, 8