SPACE-TIME GUIDED ASSOCIATION LEARNING FOR UNSUPERVISED PERSON RE-IDENTIFICATION

Chih-Wei Wu[⋆]*, Chih-Ting Liu*[⋆]*, Wei-Chih Tu*[⋆]*, Yu Tsao*[‡]*, Yu-Chiang Frank Wang*[†]*, Shao-Yi Chien*[⋆]

[⋆]Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan
[†]Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan
[‡]Research Center for IT Innovation, Academia Sinica, Taipei, Taiwan

ABSTRACT

Person re-identification (Re-ID) aims to match images of the same person across distinct camera views. In this paper, we propose the Space-Time Guided Association Learning (STGAL) for unsupervised Re-ID without ground truth identity nor image correspondence observed during training. By exploiting the spatial-temporal information presented in pedestrian data, our STGAL is able to identify positive and negative image pairs for learning Re-ID feature representations. Experiments on a variety of datasets confirm the effectiveness of our approach, which achieves promising performance when comparing to the state-of-the-art methods.

***Index Terms*—** Person re-identification, unsupervised learning, computer vision

1. INTRODUCTION

Person re-identification (Re-ID) aims to solve the problem of identifying pedestrians in a camera network. It is fundamentally challenging as pedestrians often yield different poses, scales, and lighting conditions under different camera views. While deep learning methods [1, 2, 3, 4] have demonstrated promising performance, they typically rely on labeled datasets to train their models. The task of *unsupervised person Re-ID* would be more practical yet more difficult to address. Existing approaches [5, 6, 7] focus on transferring Re-ID knowledge from a source domain (with ground truth labels) to the unlabeled target domain. However, as noted by Li *et al*. [8], these domain transfer methods typically overlook the data discrepancy between domains. Instead of selecting a "proper" source domain to transfer, some approaches [8, 9, 10] aim to estimate ID labels directly on the unlabeled domain. Yet, their performance still significantly falls behind their supervised learning counterpart.

In addition to visual structure, the *spatial-temporal information* captured along with images, i.e. time stamps and location, has been exploited to assist pedestrian identification in the literature of multi-camera tracking [11, 12, 13] and person Re-ID [14, 15]. This information enables us to understand the pedestrian "traveling pattern" between cameras. As depicted in Figure 1, the traveling pattern characterizes the time for traveling from one camera to another. Therefore, by observing the time difference between two images, it would be possible to alleviate the appearance ambiguity problem. However, the main challenges of utilizing spatial-temporal information in unlabeled settings are: 1) One could not easily establish the traveling patterns without the ground truth identity of each image. 2) It is not clear how to convert the traveling pattern information into pedestrian labels for Re-ID purposes. Existing works either rely on labeled images to build traveling patterns [11, 12, 13, 15] or require learning models with parameter-tuning efforts [14]. Thus, it is still a challenging task to realize the above idea in unlabeled settings.

To address these challenges, we propose the *Space-Time Guided Association Learning (STGAL)* framework for learning person Re-ID features with spatial-temporal information presented in the captured pedestrian images. In particular, we exploit the time stamps and the camera view of the images in our STGAL framework. As illustrated in Figure 1, we first introduce an adaptive method to construct traveling patterns from unlabeled images. On top of it, we establish a robust *Iterative Best-Buddies Search* algorithm and hard sample mining technique to predict training labels by associating positive (same ID) and negative (different ID) image pairs. During inference, we reconstruct the traveling pattern using the trained model to refine the final person Re-ID results. We note that our method does not depend on sensitive parameters and iterative refinement process to learn the model, therefore it is robust and practically preferable than previous work [14]. Our contributions are highlighted as follows:

- We propose the STGAL framework to leverage spatial-temporal information for unsupervised person Re-ID.
- We introduce an adaptive method to construct pedestrian traveling patterns without ground truth ID.
- We develop an robust matching algorithm and hard negative mining techniques to predict reliable image pairs for training Re-ID models.

This research was supported in part by the Ministry of Science and Technology of Taiwan (MOST 108-2633-E-002-001), National Taiwan University (NTU-108L104039), Intel Corporation, Delta Electronics and Compal Electronics.

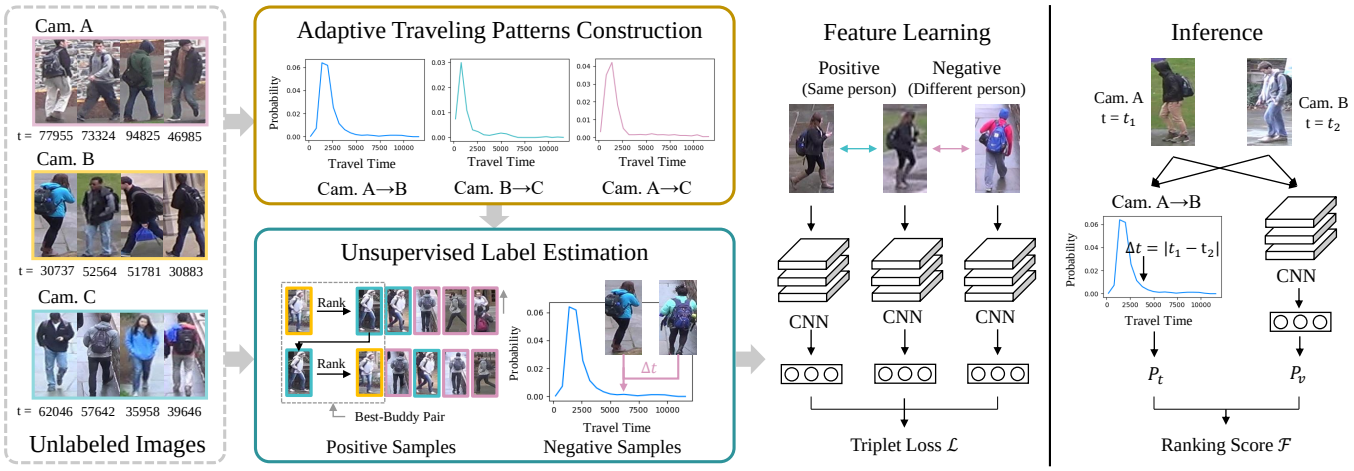

Fig. 1: **Overview of our STGAL framework.** To leverage spatial-temporal information from cross-camera pedestrian images, we develop an adaptive method for constructing pedestrian traveling patterns (Sec. 2.1), followed by a robust matching algorithm with hard sample mining techniques to predict positive and negative image pairs for unsupervised re-ID (Sec. 2.2).

2. SPACE-TIME GUIDED ASSOCIATION LEARNING

Given a set of unlabeled pedestrian images X, our goal is to learn a function $\mathcal{F}\colon (x_i, x_j) \rightarrow \mathbb{R}$ that predicts the confidence of two images $x_i, x_j \in X$ belonging to the same identity. In this paper, we focus on predicting pseudo labels (i.e., positive and negative image pairs) with spatial and temporal information. We first introduce an adaptive method to construct traveling patterns of pedestrians in Section 2.1, followed by a robust matching algorithm to perform unsupervised labeling in Section 2.2. During inference, the traveling patterns will be utilized to obtain the final Re-ID results (Section 2.3). The framework of our proposed model is illustrated in Figure 1.

2.1. Constructing Adaptive Traveling Patterns from Unlabeled Data

As shown in Figure 1, the pedestrian traveling patterns characterize the time spent to travel between any camera pair. The traveling patterns could serve as a cue to discriminate visually ambiguous pedestrians. Previous work [11, 12, 13, 15] typically construct traveling patterns with labeled training data, which is not feasible under our unsupervised setting. While Lv *et al.* [14] also propose an unsupervised approach for building traveling patterns, their method requires different sets of heuristic parameters to work under different datasets, which is not fully unsupervised essentially. In contrast, we propose an adaptive method to automatically estimate the traveling pattern without any identity labeling.

Given a feature extractor ϕ and x_i, x_j as images captured by cameras c_i, c_j, respectively, we predict their visual association probability by:

$$P_v(x_i, x_j) = e^{-\alpha \cdot D(\phi(x_i), \phi(x_j))}, \tag{1}$$

where α is a fixed scaling factor and $D(\cdot, \cdot)$ represents the Euclidean distance function. To build traveling patterns without ground truth identity, we seek guidance from the intra-camera characteristics of surveillance videos. More specifically, we utilize the fact that an image pair captured at the same time in the same camera view must be of *different* identities, i.e., a negative image pair. These negative pairs allow us to automatically determine the matching criteria in a data-oriented fashion. To observe the matched image pairs across cameras, we have:

$$M_{c_i,c_j} = \left\{ (x_i, x_j) \colon P_v(x_i, x_j) > \sigma \;\middle|\; x_i \in X_{c_i}, x_j \in X_{c_j} \right\}, \tag{2}$$

where M_{c_i,c_j} denotes the matched image pairs between camera c_i, c_j and X_{c_i}, X_{c_j} denotes the images captured by camera c_i, c_j. Note that σ is automatically determined by the largest feature distance observed in the aforementioned negative pairs through:

$$\sigma = \max\left(P_v(x_i^a, x_i^b)\right), \tag{3}$$

where (x_i^a, x_i^b) are any intra-camera negative pairs.

To further exploit the spatial-temporal information of pedestrian images, we observe their traveling patterns by constructing the PDF of traveling time between cameras c_i, c_j based on M_{c_i,c_j}:

$$P_t(\Delta | c_i, c_j) = \frac{|\{(x_i, x_j) \colon \Delta - \epsilon < |t_i - t_j| < \Delta + \epsilon\}|}{|M_{c_i,c_j}|}, \tag{4}$$

where $(x_i, x_j) \in M_{c_i,c_j}$ are matched cross-camera images with time stamps t_i, t_j and ϵ is a short time interval. The traveling probability P_t provides auxiliary information to discriminate visually ambiguous images. The visual similarity P_v and traveling probability P_t can be further combined to estimate a more robust association probability P_a of an image pair by:

$$P_a(x_i, x_j, \Delta, c_i, c_j) = P_v(x_i, x_j) \cdot P_t(\Delta | c_i, c_j). \tag{5}$$

Fig. 2: **Iterative best-buddies search algorithm.** Images in yellow bounding boxes denote the query image. Images in green have the same ID as the query, while images in purple have different IDs. Note that our searching algorithm identifies positive image pairs based on the travelling patterns observed from unlabeled pedestrian data.

2.2. Iterative Best-Buddies Search for Label Prediction

With the estimated traveling patterns, we propose algorithms discover useful positive and negative image pairs from the target domain in a fully unsupervised setting. The collected positive and negative image pairs are then used for training the feature extractor ϕ.

For discovering positive pairs, we develop a robust matching algorithm inspired by the Best-Buddies Pairs (BBP) concept in the field of template matching [16]. Given a query image, we rank all cross-camera images with the association probability P_a. A best-buddies pair in this case is a pair of images that mutually treat each other as their top-ranked candidate. The BBP approach can discover one (but only one) reliable positive match for a query. Yet, it fails to discover other valid positive matches that are important to help us learn robust feature extractor. Take three images of the same ID captured by three distinct cameras as an example. Ideally, any two images form a positive pair, so we have a total of $\binom{3}{2} = 3$ effective positive pairs. However, the standard BBP only allows an image to be matched with one another, so only one positive pair can be found. To overcome this problem, we propose the *Iterative Best-Buddies Search (IBBS)* to find multiple matches for a single query image. As depicted in Figure 2, after performing BBP once, we remove the top-ranked sample (current best-buddy) from the candidate pool and run the BBP again. By iterating the process until no more BBP is found, we can discover multiple valid positive pairs for training. In practice, we perform IBBS for all camera pairs simultaneously to discover positive samples automatically.

As for labeling negative pairs, the most straightforward way would be to regard image pairs with low P_a as negatives. Yet, we find these samples ineffective for training since most negatives could be easily distinguished by the pretrained feature extractor. Instead, we select a pool of image pairs with P_a lower than the median of the data and sample negative pairs from the pool with probability in proportion to their P_v. These samples serve as the hard negative examples to teach the feature extractor to be more robust because they are visually similar but are disagreed by the traveling patterns.

2.3. Model Learning and Inference

With the predicted (pseudo) positive and negative pairs, we train our feature extractor with the batch-hard soft-margin triplet loss [17, 2]:

$$\mathcal{L} = \sum_{q,p,n} \log(1 + e^{\|\phi(x_q)-\phi(x_p)\|_2 - \|\phi(x_q)-\phi(x_n)\|_2}). \quad (6)$$

where x_q, x_p, x_n denotes the query, positive and negative image. The triplet loss enforces the feature extractor to associate positive pairs and separate negative pairs in the embedding space. Though iterating the labeling and training process may improve the extractor, we find our predicted labels sufficiently robust to achieve optimal performance in just one-step, which is more time-efficient. During inference, we re-estimate the pedestrian traveling patterns using techniques introduced in Section 2.1, and utilize both the visual feature and the traveling pattern to calculate the final ranking function $\mathcal{F}(x_i, x_j; \phi) \equiv P_a(x_i, x_j, \Delta, c_i, c_j)$.

3. EXPERIMENTS

3.1. Experimental Settings

We conduct experiments on person Re-ID datasets that comes with time stamps and location information. All ground truth ID labels are excluded during the entire training process. Market-1501 [18] consists of 32,668 pedestrian images and 1,501 IDs captured by 6 cameras. DukeMTMC-ReID [19] contains 36,411 pedestrian images of 1,812 IDs captured by 8 cameras. In the rest of this paper, we abbreviate Market-1501, and DukeMTMC-ReID as Market and Duke. As for evaluation, we report the rank-1 accuracy (R1) of Cumulative Matching Characteristics and the Mean Average Precision (mAP) [18].

3.2. Implementation Details

We adopt ResNet-50 [20] as our feature extractor ϕ and use the 2048-d feature after the last pooling layer to represent the input image. We pretrain our feature extractor with intra-camera labels estimated by SSTT [8], an unsupervised labeling method. We note that our method also works with

Table 1: **Performances of unsupervised Re-ID methods.** Note that "Transfer" indicates use of other source domain labeled data, while "S.T." denotes observation of spatial-temporal information. For our method, we report the performance of retrieving images with feature distance P_v in Eq. 1, and with association probability P_a in Eq. 5.

Method	Supervision Category	Market		Duke	
		R1	mAP	R1	mAP
BOW [18]	N/A	35.8	14.8	17.1	8.3
PUL [10]	N/A	45.5	20.5	30.0	16.4
SPGAN [7]	Transfer	58.1	26.9	46.4	26.2
TFusion [14]	S.T.	60.8	-	-	-
TAUDL [8]	N/A	63.7	41.2	61.7	43.5
BUC [9]	N/A	66.2	38.3	47.4	27.5
ARN [6]	Transfer	70.3	39.4	60.2	33.4
ECN [5]	Transfer	75.1	43.0	63.3	40.4
Ours (P_v)	S.T.	72.1	48.4	68.4	47.1
Ours (P_a)	S.T.	**93.1**	**63.5**	**86.0**	**68.5**

other general pretrained models as demonstrated in later experiments. During training, we optimize Eq. 6 for 30,000 iterations with Stochastic Gradient Descent (SGD) with learning rate of 0.0005. In all experiments, we fix α in Eq. 1 to be 0.01 and ϵ in Eq. 4 to be 10 seconds.

3.3. Comparisons with State-of-the-arts

We compare our STGAL with existing unsupervised person Re-ID methods in Table 1. When retrieving with pure visual features (P_v in Eq. 1), our method is able to compete with state-of-the-art methods on both datasets. Note that unlike domain transfer methods, ours does not require another "relavent" labeled dataset during training. The model gains performance from reliable pseudo-labels predicted by our IBBS algorithm and hard negative examples. Furthermore, when retrieving with the help of traveling patterns (P_a in Eq. 5), which is the full version of our framework, our STGAL improves over the best performer by 18% and 22.7% in R1 on Market and Duke. This significant improvement attributes to the additional cues provided from space-time information. The traveling patterns predicted by our method rule out large proportion of improbable matches and narrow down to those that are physically plausible.

3.4. Ablation Study

The effectiveness of STGAL. In Table 2, we study the effectiveness of STGAL with different baseline models. We report the retrieval results using P_v (Eq. 1) in this table to focus on the feature extractor performance. First, we compare to a feature extractor pretrained on an unrelated Re-ID dataset following the convention of previous work [14, 10] (Baseline A). Our STGAL gains 20.6% and 16.7% in R1 on Market and Duke thanks to the reliable labels predicted with the help of traveling patterns and robust algorithms. Moreover, as our method focuses on predicting inter-camera labels for unsupervised Re-ID, we note that STGAL could be combined with any intra-camera label estimation techniques such as [8]. Therefore, we demonstrate a baseline where we pretrain our feature extractor on intra-camera labels using SSTT [8] (Baseline B). By applying STGAL on this baseline, we improve performance by 21.4% and 22.7% in R1 on Market and Duke. The consistent gain on both baselines attributes to our adaptive construction of traveling patterns and reliable postive and negative labeling methods. The performance gain on both baselines confirms the effectiveness of our method.

Table 2: **Effectiveness of STGAL.** Baseline A: Pretrain on an unrelated Re-ID dataset. Baseline B: Pretrain on estimated intra-camera labels [8].

Methods	Market		Duke	
	R1	mAP	R1	mAP
Baseline A	30.2	10.1	23.2	10.7
+STGAL	**50.8**	**20.3**	**39.9**	**16.5**
Baseline B	50.7	26.2	45.7	26.6
+STGAL	**72.1**	**48.4**	**68.4**	**47.1**

Table 3: **Effectiveness of Iterative Best-buddies Pair Algorithm.**

Method	Market		Duke	
	R1	mAP	R1	mAP
BBP	62.4	33.6	67.2	46.1
IBBS (Ours)	**72.1**	**48.4**	**68.4**	**47.1**

Effectiveness of Iterative Best-Buddies Pair algorithm. In Table 3, we further analyze the performance of our IBBS algorithm tailored for multi-camera Re-ID in comparison to the standard BBP described in Section 2.2. We report the retrieval results of P_v (Eq. 1) in this table to better visualize feature extractor performance. The feature extractor trained with IBBS labeled positive pairs gains 14.8% and 1.0% in mAP on Market and Duke datasets. Our IBBS is able to discover more valid positive pairs than the BBP method because IBBS does not constrain an image to be matched to at most one another. And by robustly extracting more positive pairs, our model is able to learn with more training samples and therefore resulting in better performance.

4. CONCLUSION

In this paper, we present a novel STGAL framework to learn Re-ID features without any ID ground truth. We generate psuedo labels for image pairs by exploiting spatial-temporal information within pedestrian images to train our feature extractor. Through the experiments, we confirm the effectiveness of our algorithm design and achieve promising performance on several unsupervised person Re-ID benchmarks.

5. REFERENCES

[1] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah, "Human semantic parsing for person re-identification," in *IEEE CVPR*, 2018, pp. 1062–1071.

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

[3] Liang Zheng, Yi Yang, and Alexander G Hauptmann, "Person re-identification: Past, present and future," *arXiv*, 2016.

[4] Wei Li, Xiatian Zhu, and Shaogang Gong, "Harmonious attention network for person re-identification," in *IEEE CVPR*, 2018, vol. 1, p. 2.

[5] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *IEEE CVPR*, 2019, pp. 598–607.

[6] Yu-Jhe Li, Fu-En Yang, Yen-Cheng Liu, Yu-Ying Yeh, Xiaofei Du, and Yu-Chiang Frank Wang, "Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification," in *IEEE CVPR Workshop*, 2018.

[7] Weijian Deng, Liang Zheng, Guoliang Kang, Yi Yang, Qixiang Ye, and Jianbin Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *IEEE CVPR*, 2018.

[8] Minxian Li, Xiatian Zhu, and Shaogang Gong, "Unsupervised person re-identification by deep learning tracklet association," in *ECCV*, 2018, pp. 737–753.

[9] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *AAAI*, 2019, vol. 33, pp. 8738–8745.

[10] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 4, pp. 83, 2018.

[11] Kuan-Wen Chen, Chih-Chuan Lai, Yi-Ping Hung, and Chu-Song Chen, "An adaptive learning method for target tracking across multiple cameras," in *IEEE CVPR*. IEEE, 2008, pp. 1–8.

[12] Cheng-Hao Kuo, Chang Huang, and Ram Nevatia, "Inter-camera association of multi-target tracks by online learned appearance affinity models," in *ECCV*. Springer, 2010, pp. 383–396.

[13] Yinghao Cai and Gerard Medioni, "Exploring context information for inter-camera multiple target tracking," in *IEEE WACV*. IEEE, 2014, pp. 761–768.

[14] Jianming Lv, Weihang Chen, Qing Li, and Can Yang, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns," in *IEEE CVPR*, June 2018.

[15] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie, "Spatial-temporal person re-identification," in *AAAI*, 2019, pp. 8933–8940.

[16] Tali Dekel, Shaul Oron, Michael Rubinstein, Shai Avidan, and William T Freeman, "Best-buddies similarity for robust template matching," in *IEEE CVPR*, 2015, pp. 2021–2029.

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE CVPR*, 2015, pp. 815–823.

[18] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *IEEE ICCV*, 2015, pp. 1116–1124.

[19] Zhedong Zheng, Liang Zheng, and Yi Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *IEEE ICCV*, 2017.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016, pp. 770–778.