End-to-End Trainable Trident Person Search Network Using Adaptive Gradient Propagation

Byeong-Ju Han, Kuhyeun Ko, and Jae-Young Sim*
Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea
{bjhan, khko, jysim}@unist.ac.kr

Abstract

Person search suffers from the conflicting objectives of commonness and uniqueness between the person detection and re-identification tasks that make the end-to-end training of person search networks difficult. In this paper, we propose a trident network for person search that performs detection, re-identification, and part classification together. We also devise a novel end-to-end training method using adaptive gradient weighting that controls the flow of back-propagated gradients through the re-identification and part classification networks according to the quality of the person detection. The proposed method not only prevents the over-fitting but encourages to exploit fine-grained features by incorporating the part classification branch into the person search framework. Experimental results on the CUHK-SYSU and PRW datasets demonstrate that the proposed method achieves the best performance among the state-of-the-art end-to-end person search methods.

1. Introduction

Person search has drawn considerable attention recently with the increasing demand for person re-identification in real world images. The conventional methods of person re-identification consider a set of cropped person images and find the person images matching to a given query person. On the other hand, the person search methods directly find the target persons from a set of scene images where multiple persons may appear in a single image. Therefore, person search has a strong potential to be applied to numerous practical applications. For example, it can be used to deploy visual surveillance systems to monitor and trace the suspects from video sequences. Also, it can be applied to augmented reality systems with mobile camera devices that provide useful visual information for social entertainment.

The person search is an integrated task of person detection and re-identification. According to the relationship between these two sub-tasks, the existing person search methods are classified into two categories: *end-to-end* framework [21, 7, 18, 22, 3, 2, 24, 16, 5, 14, 13, 1] and *two-step* framework [23, 4, 12, 9, 19, 8]. The end-to-end methods share the features for detection and re-identification and train the networks to minimize both the detection and re-identification losses simultaneously. While the person detection tries to model the common features of persons distinct from the background, the re-identification pursues unique features to recognize the identity of each person. It is, therefore, one of the main challenges in person search to jointly train the detection and re-identification networks by coordinating these two inconsistent objectives. On the other hand, the two-step methods explicitly divide the person search task into detection and re-identification, and train the corresponding two sub-networks independently. Thus, the two-step networks are relatively easy to be trained, but at the same time, suffer from the difficulty of reflecting the re-identification results to detection.

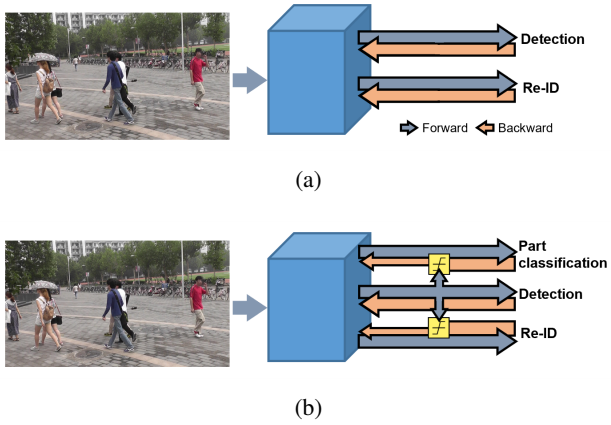


Figure 1: The concept of adaptive training for the proposed person search network. (a) Conventional end-to-end training. (b) The proposed end-to-end training.

In this paper, we propose an end-to-end person search network and its adaptive training algorithm for task-consistency between person detection and re-identification. While the objectives conflict between detection and re-identification, the re-identification performance depends

on the detection results. We devise an adaptive gradient weighting function (AGWF) that controls the flow of the back-propagated gradients through the re-identification network according to the quality of detection results. For example, when the detection network is in an early phase of training or outputs unexpected erroneous proposals, the propagation of gradients to update the parameters of re-identification network is suppressed. Moreover, we additionally employ a part classification network that refines the features of local regions in person proposals to prevent the shared features from being biased to a specific task and/or redundant training data. Figure 1 shows the proposed trident network composed of detection, re-identification, and part classification branches, and its end-to-end training where the detection confidence is reflected to the loss computation of the re-identification and part classification networks via AGWF.

The main contributions of this paper are summarized as follows.

- We proposed an end-to-end trident person search network composed of detection, re-identification, and part classification branches to generalize the shared features and alleviate the overfitting during training.
- We devised an adaptive gradient propagation scheme for end-to-end training that controls the gradient flow through the re-identification and part classification branches according to the detection confidence.
- We showed that the proposed method provides the best performance among the existing state-of-the-art end-to-end person search methods on widely used benchmark datasets of CUHK-SYSU and PRW.

2. Related Works

Two-step framework. Two-step methods train the person detection and re-identification networks separately to prevent the conflict between the two tasks. Zheng et al. [23] provided a benchmark dataset of PRW, and brought useful insights for person search through extensive experiments using the state-of-the-art person detection and re-identification networks. To learn more robust feature representation, Chen et al. [4] put more attention to the foreground regions with person mask maps obtained by a pre-trained segmentation network, and Lan et al. [12] extracted multi-scale features dealing with diverse sizes of persons in input images. Han et al. [9] refined the bounding boxes of detected persons to further contain the neighboring regions and exploit additional distinct features identifying persons from each other. Wang et al. [19] alleviated the task inconsistency between detection and re-identification by employing a mixed training set for re-identification that contains the person images cropped by the ground truth bounding boxes and the person proposals detected by a pre-trained detection network. Dong et al. [8] proposed a query-guided proposal detector that activates the features similar to the query feature such that only query-like proposals are used for re-identification.

End-to-end framework. End-to-end methods share the features to jointly train two sub-networks for person detection and re-identification. Xiao et al. [21] first provided a benchmark dataset of CUHK-SYSU, and proposed a deep learning framework that jointly trains the detection and re-identification networks. Dong et al. [7] investigated the negative effect of the background regions around each person, and designed an instance-aware branch that compares the person features extracted from a cropped person image and the entire image. Tian et al. [18] jointly optimized different tasks of instance segmentation and key-point detection to train a generalized feature representation. Yan et al. [22] applied graph convolution to extract the feature of a person region as well as the context feature of the neighboring persons. On the other hand, Chen et al. [3] alleviated the task conflict between the person detection and re-identification by using the magnitude and angle of embedded feature to obtain the detection and re-identification results, respectively. Chen et al. [2] developed a hierarchical structure of detection and re-identification by considering the sum of the probabilities that a proposal belongs to each identity as the probability of the foreground. Zhong et al. [24] aligned visible human parts together to handle the challenging cases that the persons are occluded by other objects or cut by the image boundaries. Also, attempts have been made to exploit unlabeled person data caused by incomplete annotation. Shi et al. [16] assigned the identity label to an unlabeled person whose feature is most similar to that of the labeled person. Dai et al. [5] assigned a new pseudo label to each unlabeled person using the fact that the persons appeared in the same image have different identities. Whereas most of the existing methods detect person proposals and search for the proposal best matching to the query, several methods detect query-like persons instead of all person proposals. Munjal et al. [14] localized the query person in a gallery image by allocating channel-wise weights to the features for person detection which are highly correlated to the query feature. Liu et al. [13] recursively shrank the search area of the entire scene to a local region including a query-like person. Chang et al. [1] selected the best action from a discrete action list repeatedly to transform the current bounding box to a localized query-like region in a gallery image.

In this paper, we propose a novel end-to-end training framework for person search where the multiple tasks of person detection, re-identification, and the additional third-party task of part-classification are systematically coupled together to learn representative features and avoid the task conflict between person detection and re-identification.

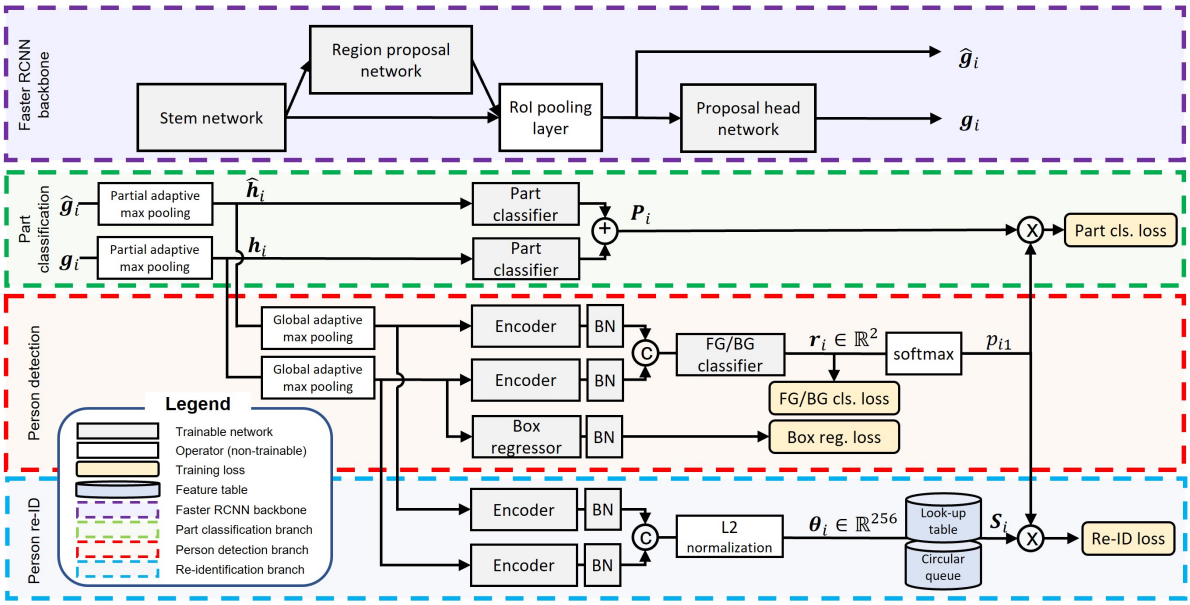


Figure 2: The overall architecture of the proposed network.

3. Classification Network Training Revisited

The cross-entropy loss has been widely used to train deep neural networks for classification. Let $\boldsymbol{S}_i$ denote the score vector of a given i-th query, where its j-th element S_{ij} represents the score that the i-th query belongs to the j-th class. The cross-entropy loss with the softmax probabilities is defined as

$$\mathcal{L}(\boldsymbol{S}_i, y_i) = -\log\left(\frac{\exp(S_{iy_i}/\tau)}{\sum_j \exp(S_{ij}/\tau)}\right), \tag{1}$$

where y_i denotes the true class index of the i-th query and τ is a temperature coefficient which controls the sensitivity of the loss function. Eq. (1) can be re-written as

$$\mathcal{L}(\boldsymbol{S}_i, y_i) = \log\left(1 + \sum_{j \neq y_i} \exp\left((S_{ij} - S_{iy_i})/\tau\right)\right), \tag{2}$$

where we see that the classification loss actually depends on the relative scores of the classes with respect to that of the true class. The loss function returns a high value when the scores that the input query is predicted to belong to incorrect classes are relatively higher than that of the true class, *i.e.* $S_{ij} > S_{iy_i}, \forall j \neq y_i$.

During training, the network parameters are updated by back-propagated gradients based on the chain rule given by

$$\frac{\partial \mathcal{L}(\boldsymbol{S}_i, y_i)}{\partial \boldsymbol{\pi}} = \sum_k \frac{\partial \mathcal{L}(\boldsymbol{S}_i, y_i)}{\partial S_{ik}} \cdot \frac{\partial S_{ik}}{\partial \boldsymbol{\pi}}, \tag{3}$$

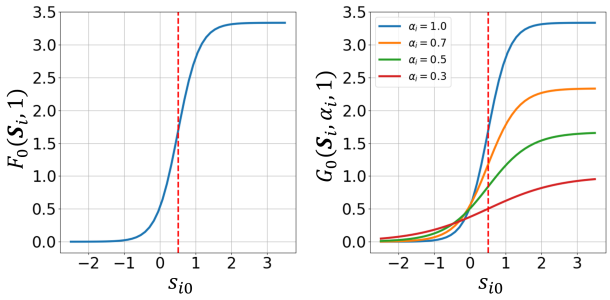
where $\boldsymbol{\pi}$ denotes the network parameters. Let $F_k(\boldsymbol{S}_i, y_i)$ be the derivative of the loss function with respect to S_{ik}, then we regard $F_k(\boldsymbol{S}_i, y_i)$ as a weighting function of the gradient $\frac{\partial S_{ik}}{\partial \boldsymbol{\pi}}$ associated with the network parameters. To interpret the behavior of this weighting function, $F_k(\boldsymbol{S}_i, y_i)$ with the loss function in (2) is derived as a softmax-like function of the relative scores given by

$$\begin{aligned} F_k(\boldsymbol{S}_i, y_i) &= \frac{\partial \mathcal{L}(\boldsymbol{S}_i, y_i)}{\partial S_{ik}} \\ &= \begin{cases} \frac{1}{\tau} \frac{\exp\left((S_{ik} - S_{iy_i})/\tau\right)}{1 + \sum_{j \neq y_i} \exp\left((S_{ij} - S_{iy_i})/\tau\right)}, & \text{if } k \neq y_i, \\ -\sum_{j \neq y_i} F_j(\boldsymbol{S}_i, y_i), & \text{otherwise.} \end{cases} \end{aligned} \tag{4}$$

For intuitive understanding, we visualize the shape of the weighting function $F_0(\boldsymbol{S}_i, 1)$ by taking an example of simple binary classification task where the true and incorrect class indices are 1 and 0, respectively. Figure 3(a) plots $F_0(\boldsymbol{S}_i, 1)$ according to varying S_{i0} when fixing $S_{i1} = 0.5$ and $\tau = 0.3$, where we see that it exhibits a logistic curve of the relative scores. It means that, when S_{i0} is relatively higher than S_{i1}, the network parameters are significantly deviated during training by propagating the magnitude of the gradient $\frac{\partial S_{ik}}{\partial \boldsymbol{\pi}}$ to the maximum with a large weight close to $\frac{1}{\tau}$. On the contrary, when S_{i0} is lower than S_{i1}, a small weight close to 0 suppresses the gradient back-propagation.

4. Proposed Method

The proposed method extracts the initial features from input images using Faster RCNN [15], which are then shared by the three branches of person detection, re-identification, and part classification. Figure 2 shows the overall structure of the proposed trident network where the



(a) (b)

Figure 3: Gradient weighting functions for the binary classification task as varying S_{i0} with fixed $S_{i1} = 0.5$ and $\tau = 0.3$. (a) The standard form of $F_0(\boldsymbol{S}_i, 1)$. (b) The proposed adaptive weighting function of $G_0(\boldsymbol{S}_i, \alpha_i, 1)$ with four different values of α_i.

person detection, re-identification, and part classification branches are highlighted in red, blue, and green, respectively. The proposed method uses both the input $\hat{\boldsymbol{g}}_i$ and output $\boldsymbol{g}_i$ of the proposal head network as features based on multi-scale featuring. As an end-to-end method, the person detection, re-identification, and part classification networks are trained simultaneously, where we use the confidence of detection to control the gradient back-propagation adaptively through the re-identification and part classification networks.

4.1. Adaptive Gradient Back-Propagation

Person search is one of the classification tasks which aims to classify the detected person proposals into their true identities. In practice, the score S_{ij} is usually computed as the cosine similarity between the features of the i-th proposal and the j-th identity. The re-identification loss is designed to train the deep networks such that the feature of the i-th proposal to be more similar to that of the true matching identity compared with that of the other identities.

Note that both person detection and re-identification are performed for person search, and therefore the re-identification features highly depend on the quality of the estimated person proposals. In particular, the end-to-end methods share the feature maps for the two tasks, and the representativeness of re-identification features is often degraded when the feature maps of the input image are overfit to person detection. We may alleviate such a drawback by training the re-identification network with a pre-trained person detection network, however, the pre-training requires heuristic decisions. Also, we may adopt the two-stage approach that uses independent feature maps for person detection and re-identification, respectively, but it suffers from high computational complexity and cumbersome training.

We propose an end-to-end training method for person search by training the re-identification network adaptively according to the confidence of detected person proposals. Regarding $F_k(\boldsymbol{S}_i, y_i)$ in (4) as a standard weighting function that controls the flow of the gradients when training deep neural networks, we devise the AGWF that reflects a proper amount of gradients to update the network parameters according to the states of training. Specifically, AGWF is designed as

$$G_k(\boldsymbol{S}_i, \alpha_i, y_i) = \alpha_i \cdot F_k(\alpha_i \cdot \boldsymbol{S}_i, y_i), \tag{5}$$

where α_i denotes the detection confidence of the i-th proposal that is set to the probability for the i-th proposal to belong to the person class.

Figure 3(b) shows the shapes of AGWF $G_0(\boldsymbol{S}_i, \alpha_i, 1)$ in the binary classification task with four different values of α_i, where we see that both the output range and the sensitivity to the input are adjusted compared to the standard form in Figure 3(a). As α_i decreases, the output range of AGWF decreases and the shape of AGWF becomes smooth. It is worth to note that AGWF implicitly reflects the current training state of the shared feature maps as well as the quality of the box regression. For example, when the network is in an early phase of training and the detected person proposal contains a significant portion of the background, the confidence α_i becomes low and therefore AGWF becomes insensitive to the feature similarities of the proposal to the identity classes while suppressing the backward gradient flow from the re-identification loss.

4.2. Re-identification

Similar to [21], the proposed method employs a look-up table that stores the representative features of identities and a circular queue that temporarily contains the features extracted from the proposals without the matching identity labels. The proposed re-identification loss $\mathcal{L}_{\text{re-id}}$ is composed of $\mathcal{L}_{\text{id}}$ and $\mathcal{L}_{\text{bin}}$ for labeled proposals and $\mathcal{L}_{\text{un}}$ for unlabeled proposals defined as

$$\begin{aligned}\mathcal{L}_{\text{re-id}}(\boldsymbol{S}_i, \alpha_i, y_i) = &\frac{1}{|\Psi_L|}\sum_{i \in \Psi_L}\{\mathcal{L}_{\text{id}}(\boldsymbol{S}_i, \alpha_i, y_i) + \mathcal{L}_{\text{bin}}(\boldsymbol{S}_i, \alpha_i, y_i)\} \\ &+ \frac{\lambda}{|\Psi_U|}\sum_{i \in \Psi_U}\mathcal{L}_{\text{un}}(\boldsymbol{S}_i, \alpha_i),\end{aligned} \tag{6}$$

where Ψ_L and Ψ_U denote the index sets of the labeled proposals and the unlabeled proposals in a mini-batch, respectively, and λ is experimentally set to 0.1. We set the confidence α_i as the probability p_{i1} that the i-th proposal belongs to the person class as shown in Figure 2.

Classification loss for labeled data. We obtain the revised loss function $\mathcal{L}_{\text{id}}$ such that its derivative with respect to S_{ik} becomes $G_k(\boldsymbol{S}_i, \alpha_i, y_i)$ in (5). Comparing to the vanilla

classification loss in (2) and its derivative in (3), we have

$$\mathcal{L}_{\text{id}}(\boldsymbol{S}_i, \alpha_i, y_i) = \log\left(1 + \sum_{\substack{k\in\{\Omega_L\cup\Omega_Q\}\\ k\neq y_i}} \exp(\alpha_i(S_{ik} - S_{iy_i})/\tau_1)\right), \tag{7}$$

where Ω_L and Ω_Q represent the index sets of the look-up table and the circular queue, respectively. Note that the classification loss depends on the relative similarity, and S_{iy_i} itself is often saturated before reaching a high score. Thus we additionally employ a binary classification loss $\mathcal{L}_{\text{bin}}$ to encourage S_{iy_i} to become high.

$$\mathcal{L}_{\text{bin}}(\boldsymbol{S}_i, \alpha_i, y_i) = \log\left(1 + \exp((s_\kappa - \alpha_i S_{iy_i})/\tau_2)\right), \tag{8}$$

where the constant s_κ is set to the maximum similarity of 1. In Figure 2, the skip connection between the person detection branch and the re-identification branch represents the application of AGWF to the re-identification loss, where the AGWF increases the contribution of the re-identification loss to update the network as the detection results are reliable.

Classification loss for unlabeled data. Most of the existing methods utilize the unlabeled proposals, associated with the ground truth bounding boxes without identity labels, as negative samples to be compared to the labeled proposals. The proposed method explicitly computes the classification loss for the unlabeled proposals. Since the representative identity feature of an unlabeled proposal is unknown, we cannot compute S_{iy_i} in (7) for the unlabeled proposal. Instead, we simply take the average similarity of the labeled proposals as the similarity of matching identity features for the unlabeled proposals, and compute $\mathcal{L}_{\text{un}}$ as

$$\mathcal{L}_{\text{un}}(\boldsymbol{S}_i, \alpha_i) = \log\left(1 + \sum_{k\in\Omega_L} \exp(\alpha_i(S_{ik} - \bar{S})/\tau_1)\right), \tag{9}$$

where

$$\bar{S} = \frac{1}{|\Psi_L|}\sum_{i\in\Psi_L} S_{iy_i}. \tag{10}$$

Note that $\mathcal{L}_{\text{un}}$ implicitly strikes a balance between the classification abilities for the labeled proposals and the unlabeled proposals, since it depends on $\bar{S}$ that is changeable during the network training.

Distinction from related work. CWS [23] and NAE [3] compute the similarity weighted by detection confidence to refine the matching results of re-identification, however, they do not consider the detection confidence to train the person search networks. On the contrary, the proposed method uses the weighted similarity to perform re-identification as well as to compute the loss functions of re-identification and part classification in order to propagate the gradients adaptively when training the deep network. Although TCTS [19] utilizes the foreground score to train the re-identification network for the purpose of extracting features robust to erroneous detection results, it is relatively cumbersome to train the detection and re-identification networks separately. On the other hand, the proposed method adopts the detection confidence to encourage the network to extract informative features for detection as well as re-identification by training the entire networks simultaneously based on the end-to-end framework.

(a) (b)

Figure 4: Part classification. (a) Partition of person proposals. (b) Classified part images.

4.3. Part Classification

Numerous methods for person search obtain an embedded feature vector $\boldsymbol{x}_i$ by squeezing $\boldsymbol{g}_i$ in the spatial dimension. In such a case, the results of person detection or re-identification are often dominated by a certain local region of the proposal, which may prevent $\boldsymbol{g}_i$ from being updated to exhibit informative features of the other regions. The existing methods of CTX [22] and APNet [24] extract partition-based features which are used only for the re-identification task. However, as shown in Figure 2, we additionally construct a sub-network for the part classification task, and share the feature maps obtained by ResNet-50 among the three branches of part classification, person detection, and re-identification together. The proposed trident network is end-to-end trained based on the multi-task learning framework to embed distinct local features of persons to recognize person identities while utilizing more generalized features to prevent the network from being trained to overfit to a certain dataset and/or sub-task.

In practice, similar to PCB [17], we horizontally divide a cropped input image into several partitions that represent different visual structures from each other. Then we apply a part classifier to make the partition features distinct from each other as shown in Figure 4. Note that although each partition seems to represent a certain human body part, e.g., head or legs, it is not restricted to such body parts but may indicate an arbitrary region hard to be named as one of the body parts. The proposed method performs the max pooling to $\boldsymbol{g}_i$ to get a feature matrix $\boldsymbol{h}_i \in \mathbb{R}^{N_p\times 2048}$, where N_p denotes the number of partitions and $\boldsymbol{h}_{ij}$, the j-th row

of $\boldsymbol{h}_i$, represents the feature of the j-th partition of the i-th proposal. We can simply annotate the ground truth part label as $z_{ij} = j$. In addition, we also assign the dummy class label of $N_p + 1$ to the partitions for the background proposals. The part classifier composed of a fully connected layer takes the partition feature vectors $\boldsymbol{h}_{ij}$'s and outputs a prediction score matrix $\boldsymbol{P}_i \in \mathbb{R}^{N_p \times (N_p+1)}$ where its (j,k)-th element P_{ijk} represents the probability that $\boldsymbol{h}_{ij}$ belongs to the k-th partition. The classifier is trained by the cross-entropy based classification loss. We also apply AGWF to the part classification loss $\mathcal{L}_{\text{part}}$ to adaptively train the network, since the partitions depend on the quality of person detection as well.

$$\mathcal{L}_{\text{part}} = \frac{1}{N} \sum_{\substack{i \in \Psi \\ 1 \le j \le N_p}} \log\left(1 + \sum_{k \neq z_{ij}} \exp(\alpha_i (P_{ijk} - P_{ijz_{ij}})/\tau_3)\right), \tag{11}$$

where Ψ denotes the set of all proposals including the background proposals, and N means the number of total partitions in a mini-batch, i.e., $N = N_p \times |\Psi|$.

5. Experimental Results

We evaluate the performance of the proposed method on two widely used datasets of CUHK-SYSU [21] and PRW [23] compared with 18 state-of-the-art methods of person search.

5.1. Datasets and Evaluation Protocol

Benchmark datasets. CUHK-SYSU [21] dataset contains video frames sampled from movies and city scene images captured by a moving camera, where Figure 5(a) shows some example images. CUHK-SYSU provides 18,184 unconstrained images, 96,143 person bounding boxes, and 8,432 identity labels. The dataset is divided into the training set of 11,206 images with 5,532 identities and the test set of 6,978 images with 2,900 query persons. Instead of using all the test images as a gallery, it provides the lists of images for pre-defined galleries of different sizes from 50 to 4000.

PRW [23] dataset contains sampled frames from the videos captured at 6 different camera positions as shown in Figure 5(b). PRW provides 11,816 images, 43,110 bounding boxes, and 932 identity labels. The dataset is divided into the training set of 5,704 images with 482 identities and the test set of 6,112 images with 2,057 query persons. Unlike CUHK-SYSU, PRW dataset yields several variants of gallery formation. To compensate the performance variation according to different gallery formations of the existing methods, we test the proposed method on two types of galleries: *regular gallery* and *multi-view gallery*. The regular gallery denotes the gallery described in [18], which contains all test images except the query image and ignores the unlabeled persons in the test images. On the other hand, the multi-view gallery excludes the redundant images captured at the camera position of a given query image and allows unlabeled persons to be explored. The person search on the multi-view gallery is more challenging since it contains the unlabeled persons while including a relatively small number of persons matched to the query.

(a) CUHK-SYSU (b) PRW

Figure 5: Sample images in the benchmark datasets for person search. (a) CUHK-SYSU [21]. (b) PRW [23].

Evaluation protocol. The performance of person search is usually evaluated by the mean Average Precision (mAP) and Top-k scores, where mAP measures the area under the precision-recall curve for the matching results and Top-k score represents the percentage that at least one of the k proposals most similar to a given query succeeds in the identity matching. The proposals whose intersection over union (IoU) to the ground truth bounding boxes is less than 0.5 are not counted in Top-k scores.

5.2. Implementation Details

The proposed network basically uses ResNet-50 [11] pre-trained on ImageNet [6] to extract the initial features. The sequence of blocks from 'conv1' to 'conv4' in ResNet-50 are used as a stem network which extracts the spatial features from an input image, and the 'conv5' block is used as a proposal head network that embeds deep features for each proposal. We apply RoIAlign layer [10] for RoI pooling. As shown in Figure 2, the three task branches of the proposed network are composed of the sub-networks such as part classifiers, foreground classifiers, box regressor, and encoders, where each sub-network is simply defined as a fully connected layer and only the foreground classifiers do not train the biases. Each encoder outputs embedded features of the channel size 128 and the other sub-networks output the features of the proper channel sizes.

We set the hyper-parameters in the region proposal network (RPN) using [3] except that the output size of RoI pooling is set to 24×8. We initialize the weights and biases of the sub-networks in the task branches using the values drawn from the normal distribution with the standard deviation of 0.01 and the constant of 0, respectively. All the classification loss terms in the task branches have an equal

Detection				
Method	AP	Recall	Δ AP	Δ Recall
Proposed	94.54	97.49		
w/o Part cls.	93.22	97.02	-1.32	-0.47
w/o AGWF	92.30	95.20	-2.24	-2.29
w/o AGWF & Part cls.	92.67	96.36	-1.87	-1.13
Re-identification				
Method	mAP	Top-1	Δ mAP	Δ Top-1
Proposed	48.04	73.21		
w/o Part cls.	46.52	72.19	-1.52	-1.02
w/o AGWF	46.87	72.92	-1.17	-0.29
w/o AGWF & Part cls.	45.82	70.24	-2.22	-2.97

Table 1: Effect of AGWF and part classification. The performance of person detection and re-identification are evaluated on the multi-view gallery in PRW dataset.

weight of 1, and the temperature coefficients are experimentally set as $\tau_1 = 1/30$, $\tau_2 = 2$ and $\tau_3 = 1/10$. We empirically determine the sizes of the circular queue as 5000 and 500 for CUHK-SYSU and PRW datasets, respectively.

For training, we employ a mini-batch of size 4. By applying a warm-up starter and a scheduler, the learning rate linearly increases from 0 to 3×10^{-3} during the first epoch and decays by 0.1 after 8 epochs. The training continues over 14 epochs and 10 epochs for CUHK-SYSU and PRW datasets, respectively. All trainable parameters are updated by an SGD optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4} in an end-to-end learning manner without any heuristic training plan about the training order and time for each module. The look-up table and the circular queue are updated as [21] except that the circular queue temporally contains the background proposals as well.

5.3. Ablation Study

We first perform the ablation study for the proposed method tested on the PRW dataset with multi-view gallery to show the effectiveness of part classification, AGWF, and revised re-identification loss, respectively.

Effect of part classification. During training, the proposed method performs the part classification task to make the shared feature maps more descriptive. Table 1 analyzes the effect of the part classification task on the detection task and the re-identification task, respectively. The detection performance is measured in terms of AP and recall scores, and the re-identification performance is measured by mAP and Top-1 scores. We see that it improves the performance by 1.32% and 0.47% in terms of AP and recall scores for person detection and 1.52% and 1.02% in terms of mAP and Top-1 scores for re-identification, respectively. Also, each column in Figure 6 shows the tuple of a query and the two matching results with the highest similarity with the query estimated by the proposed method and a comparative method without using the part classification. We see that the proposed method usually finds the target persons correctly, who exhibit distinct local features such as the cloth patterns and accessories, e.g., bags and umbrellas. However, without the part classification, incorrect persons with similar cloth colors to the queries are returned.

Figure 6: Qualitative comparison results with and without part classification. The results of true and false matching are depicted in green and red, respectively.

Effect of adaptive gradient weighting. AGWF updates the network parameters by propagating reliable gradients adaptively. Table 1 shows AGWF is effective for both the detection and re-identification tasks. Deactivating AGWF decreases the detection performance by 2.24% and 2.29% in terms of AP and recall scores, and the re-identification performance by 1.17% and 0.29% in terms of mAP and Top-1 scores, respectively. Furthermore, without the part classification, deactivating AGWF further degrades the re-identification performance and eventually yields the worst performance scores. However, the increased number of tasks in the end-to-end framework may cause the risk of task conflict. Thus, when the part-classification is applied without AGWF, the detection performance becomes the worst even lower than that of the proposed method without both the part-classification and AGWF. Note that the AGWF is applicable for not only the proposed trident network but also conventional two task networks of person detection and re-identification.

Effect of re-identification loss. $\mathcal{L}_{\text{bin}}$ aims to increase the similarity to the positive class regardless of the similarities to the negative classes, while $\mathcal{L}_{\text{un}}$ only focuses on decreasing the similarities to the negative classes due to the unknown positive class. Therefore, when evaluating their efficacy, it would be more reasonable to use both $\mathcal{L}_{\text{bin}}$ and $\mathcal{L}_{\text{un}}$ together as complementary terms to each other to avoid undesired bias. As shown in Table 2, the method without $\mathcal{L}_{\text{bin}}$ and $\mathcal{L}_{\text{un}}$ causes performance degradation by 0.6% and 1.36% in terms of mAP and Top-1 scores, respectively.

Method	mAP	Top-1	Δ mAP	Δ Top-1
Proposed	48.04	73.21		
w/o $\mathcal{L}_{\text{bin}}$&$\mathcal{L}_{\text{un}}$	47.44	71.85	-0.60	-1.36

Table 2: Effect of the re-identification loss terms. The re-identification performance is evaluated on PRW dataset.

5.4. Comparison with State-of-the-Arts

Table 3 compares the quantitative performance of the proposed method with that of 18 existing state-of-the-art person search methods: 12 end-to-end methods and 6 two-step methods, in terms of mAP and Top-1 scores.

Performance on CUHK-SYSU dataset. The performance of the proposed method on CUHK-SYSU dataset is 93.3% and 94.2% in terms of mAP and Top-1 scores, respectively, which are the best scores among the 12 end-to-end methods and also comparable to the best scores of the two-step methods. Considering the efficiency and complexity of the end-to-end network training, the proposed method is a promising tool for person search.

Performance on PRW dataset. Due to the lack of concrete criterion for gallery formation, we evaluate the performance of the proposed method on PRW dataset according to the two different gallery formations, regular gallery and multi-view gallery, respectively. The second group of data denoted by * in Table 3 compares the performance evaluated on the multi-view gallery, and the other groups of data compare the performance on the regular gallery. The performance scores with the regular gallery are compared with that of the existing methods reported in the literatures. As shown in Table 3, the proposed method achieves the best mAP score of 53.3% among all the compared methods including the two-step methods which is significantly better than the second best score of 48.5%. Also, the proposed method achieves the second best Top-1 score of 87.7% slightly lower than the best score by 0.2%. In addition, we also re-evaluate the performance of two recent person search methods of NAE [3] and HOIM [2] on the multi-view gallery by using the pre-trained models publicly available. The multi-view gallery is a relatively challenging dataset for person search, and therefore we see that the associated performance of the compared methods becomes lower than that on the regular gallery. However, the proposed method largely outperforms NAE and HOIM in terms of both mAP and Top-1 scores even with the multi-view gallery. Note that PRW dataset contains a small number of training images but a large number of test images compared to CUHK-SYSU dataset. Therefore, all the compared methods suffer from degraded performance on PRW dataset. However, the proposed method significantly outperforms the other methods especially on PRW dataset, which means that the proposed method is able to train the representative features reliably even with a small dataset.

	Method	CUHK-SYSU		PRW	
		mAP	Top-1	mAP	Top-1
Two-step	DPM+IDE [23]	-	-	20.5	48.3
	CNN+MGTS [4]	83.0	83.7	32.6	72.1
	CNN+CLSA [12]	87.2	88.5	38.7	65.0
	FPN+RDLR [9]	93.0	94.2	42.9	70.2
	TCTS [19]	**93.9**	**95.1**	46.8	**87.5**
	IGPN+PCB [8]	90.3	91.4	**47.2**	87.0
End-to-end	OIM [20]	75.5	78.7	21.3	49.9
	IAN [20]	76.3	80.1	23.0	61.9
	NPSM [13]	77.9	81.2	24.2	53.1
	RCAA [1]	79.3	81.3	-	-
	CTX [22]	84.1	86.5	33.4	73.6
	QEEPS [14]	88.9	89.1	37.1	76.7
	BINet [7]	90.0	90.7	45.3	81.7
	PBNet [18]	90.5	88.4	48.5	**87.9**
	DIOIM [5]	88.7	89.6	36.0	76.1
	APNet [24]	88.9	89.3	41.2	81.4
	NAE [3]	92.1	92.9	44.0	81.1
	HOIM [2]	89.7	90.8	39.8	80.4
	Proposed	**93.3**	**94.2**	**53.3**	87.7
	NAE* [3]	-	-	40.0	67.5
	HOIM* [2]	-	-	36.5	65.0
	Proposed*	-	-	**48.0**	**73.2**

Table 3: Comparison of the quantitative performance in terms of mAP and Top-1 scores evaluated on CUHK-SYSU and PRW datasets. The highest scores in each group are highlighted in bold.

6. Conclusion

We proposed an end-to-end person search network composed of the three branches of detection, re-identification, and part classification. In order to alleviate the task inconsistency between detection and re-identification during training, we applied the adaptive gradient weighting scheme that reflects the confidence of detection to compute the loss functions of re-identification and part classification. We also exploited more generalized features in a multi-task learning manner by incorporating the part classification network additionally. Experimental results demonstrated that the proposed method outperforms 18 state-of-the-art person search methods in terms of mAP on PRW dataset.

Acknowledgements

This work was supported by NRF of Korea within the Ministry of Science and ICT (MSIT) under Grant 2020R1A2B5B01002725, and by Institute of Information & communications Technology Planning & Evaluation (IITP) through MSIT under Grant 20200013360011001 (Artificial Intelligence graduate school support (UNIST)) and 20170006670021001.

References

[1] Xiaojun Chang, Po-Yao Huang, Yi-Dong Shen, Xiaodan Liang, Yi Yang, and Alexander G. Hauptmann. Rcaa: Relational context-aware agents for person search. In *Proceedings of the European Conference on Computer Vision*, 2018.

[2] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Bernt Schiele. Hierarchical online instance matching for person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[3] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. Norm-aware embedding for efficient person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[4] Di Chen, Shanshan Zhang, Jian Yang, and Ying Tai. Person search via a mask-guided two-stream cnn model. In *Proceedings of the European Conference on Computer Vision*, 2018.

[5] Ju Dai, Pingping Zhang, Huchuan Lu, and Hongyu Wang. Dynamic imposter based online instance matching for person search. *Pattern Recognition*, 100:107120, 2020.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[7] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Bi-directional interaction network for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[8] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Instance guided proposal network for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[9] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. Re-id driven localization refinement for person search. In *Proceedings of the IEEE Conference on International Conference on Computer Vision*, 2019.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE Conference on International Conference on Computer Vision*, 2017.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[12] Xu Lan, Xiatian Zhu, and Shaogang Gong. Person search by multi-scale matching. In *Proceedings of the European Conference on Computer Vision*, 2018.

[13] Hao Liu, Jiashi Feng, Zequn Jie, Karlekar Jayashree, Bo Zhao, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. Neural person search machines. In *Proceedings of the IEEE Conference on International Conference on Computer Vision*, 2017.

[14] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-guided end-to-end person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.

[16] Wei Shi, Hong Liu, Fanyang Meng, and Weipeng Huang. Instance enhancing loss: Deep identity-sensitive feature embedding for person search. In *Proceedings of the IEEE Conference on International Conference on Image Processing*, 2018.

[17] Yifan Sun, Liang Zheng, Yi Yang, and Qi Tian. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision*, 2018.

[18] Kun Tian, Houjing Huang, Yun Ye, Shiyu Li, Jinbin Lin, and Guan Huang. End-to-end thorough body perception for person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[19] Cheng Wang, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. Tcts: A task-consistent two-stage framework for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[20] Jimin Xiao, Yanchun Xie, Tammam Tillo, Kaizhu Huang, Yunchao Wei, and Jiashi Feng. Ian: the individual aggregation network for person search. *Pattern Recognition*, 87:332–340, 2019.

[21] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[22] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. Learning context graph for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[23] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[24] Yingji Zhong, Xiaoyu Wang, and Shiliang Zhang. Robust partial matching for person search in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.