FastPR: One-stage Semantic Person Retrieval via Self-supervised Learning

Meng Sun
Tsinghua University
Beijing, China
sunm20@mails.tsinghua.edu.cn

Ju Ren*
Tsinghua University
Beijing, China
renju@mail.tsinghua.edu.cn

Xin Wang
Tsinghua University
Beijing, China
xin_wang@tsinghua.edu.cn

Wenwu Zhu
Tsinghua University
Beijing, China
wwzhu@tsinghua.edu.cn

Yaoxue Zhang
Tsinghua University
Beijing, China
zhangyx@mail.tsinghua.edu.cn

ABSTRACT

Semantic person retrieval aims to locate a specific person in an image with the query of semantic descriptions, which has shown great significance in surveillance and security applications. Prior arts commonly adopt a two-stage method that first extracts the persons with a pretrained detector and then finds the target matching the descriptions optimally.However, existing works suffer from high computational complexity and low recall rate caused by error accumulation in the two-stage inference. To solve the problems, we propose FastPR, a one-stage semantic person retrieval method via self-supervised learning, to optimize the person localization and semantic retrieval simultaneously. Specifically, we propose a dynamic visual-semantic alignment mechanism which utilizes grid-based attention to fuse the cross-modal features, and employs a label prediction proxy task to constrain the attention process. To tackle the challenges that real-world surveillance images may suffer from low-resolution and occlusion, and the target persons may be within a crowd,we further propose a dual-granularity person localization module through designing an upsampling reconstruction proxy task to enhance the local feature of the target person in the fused features, followed by a tailored offset prediction proxy task to make the localization network capable of accurately identifying and distinguishing the target person in a crowd. Experimental results demonstrate that FastPR achieves the best retrieval accuracy compared to the state-of-the-art baseline methods, with over 15 times inference time reduction.

CCS CONCEPTS

• **Information systems** → **Image search**; • **Computing methodologies** → **Object identification**.

KEYWORDS

Person retrieval, one-stage, cross-modal feature fusion, attention, self-supervised learning.

*Corresponding author

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10.
https://doi.org/10.1145/3503161.3548274

ACM Reference Format:
Meng Sun, Ju Ren, Xin Wang, Wenwu Zhu, and Yaoxue Zhang. 2022. FastPR: One-stage Semantic Person Retrieval via Self-supervised Learning. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), Oct. 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548274

1 INTRODUCTION

Nowadays, surveillance cameras have been deployed in every corner of cities for the purpose of public security. They are constantly producing extensive monitoring data, which not only provides sufficient data sources but also poses great challenges on processing efficiency for modern surveillance applications. Semantic person retrieval is a typical application, which aims to retrieve a person in an image using some semantic descriptions. The descriptions can be semantic labels, such as clothing color/types, accessories, age and skin color, etc., or free-form natural language descriptions of a person. It offers a convenient way for person retrieval and shows great potentials in advancing public security applications, e.g., criminal suspect searching. To achieve semantic person retrieval in a surveillance image with more complex background information than semantic descriptions, most of prior studies adopt a two-stage approach like [23, 27, 31]. Specifically, they divide the retrieval task into two sub-tasks: detecting all potential bounding box of the persons in the image and then selecting the most likely one by a ranking algorithm according to the semantic descriptions. Such two-stage models are easy to train due to the separate optimization processes, but incur high computational complexity in inference. Moreover, since the retrieval accuracy of the second stage highly depends on the detection accuracy of the first stage, two-stage models are more likely to bring error accumulation, resulting in a low recall rate during inference. This is an inevitable problem, because the images taken in real surveillance scenes usually suffer from low resolution or occlusion among the entities, causing great difficulty in identifying people accurately and comprehensively. Fig 1 shows the difference between one-stage and two-stage semantic person retrieval methods. There are also many works [1, 17, 21, 34] using the cropped images with a single person as the inputs of their retrieval models, which are lack of background information and the relation with other entities. Since the cropped images can't acquire from the surveillance images directly, these methods face the same problem of the two-stage models in handling surveillance images.

Figure 1: Comparison of one-stage and two-stage methods of semantic person retrieval.

Thus, it calls for a sophisticated one-stage approach to achieve more efficient and accurate semantic person retrieval.

Despite the significance, designing a one-stage semantic person retrieval approach is highly non-trivial under the following challenges. Since surveillance images are commonly with full background and abundant noise, there are more features contained by the images than the semantic descriptions. To solve the cross-modal feature fusion problem, previous works like[2, 7, 16] choose to make a common embedding layer and force the visual and semantic features to be closer by euclidean distance or cosine similarity in a certain dimension, while the forced similarity is ill-considered in the case of the unbalance information between real-world images and semantic descriptions. Other studies[13, 26, 28, 29, 33] use attention mechanisms to align the image and semantic features. However, the attention mechanisms usually lack full supervision and are hard to be optimized under the unequal amount of information between the visual and semantic features. Moreover, since there is a natural gap between visual and semantic features, how to perform bounding box regression in images according to semantic descriptions is another challenge for the one-stage method design. In addition, images in real surveillance scenarios usually suffer from low resolution and high noise, etc., which makes cross-modal bounding box regression more difficult.

To solve the above challenges, in this paper we propose FastPR, a one-stage semantic person retrieval method to locate the target person in real-world surveillance images. Specifically, we divide the image into $S * S$ grids, where one grid cell is responsible for the target person if the center of the person falls into this gird. We employ a *dynamic visual-semantic alignment* mechanism through the grid-based attention to match the semantic description with visual features, followed by designing a label prediction proxy task to improve the performance the cross-modal feature fusion. To address the low-resolution problem in surveillance images, we propose an upsampling reconstruction proxy task to enhance the local feature of the target person. Moreover, an offset prediction proxy task is designed for person localization, which can enable the localization network to identify and distinguish the target person in the crowd. During the localization process, we also predict the confidence of whether a grid covers the target person as a coarse-grained retrieval, which can guide the fine-grained retrieval (bounding box, exactly). These three components together constitute the *dual-granularity person localization* module in FastPR to guarantee the accuracy of person retrieval.

The main contributions are as follows:

- To the best of our knowledge, FastPR is the first one-stage semantic person retrieval approach to directly retrieve target persons from surveillance images.
- We propose a dynamic visual-semantic alignment mechanism by utilizing grid-based attention to fuse the cross-modal features and designing a label prediction proxy task to improve the attention process.
- We design a dual-granularity person localization module to precisely locate the target person, where an unsampling reconstruction proxy task is designed to handle the low resolution and occlusion challenge for real-world images, and a target-centric offset prediction task is designed to help accurately identifying the target person in the crowd.
- Extensive experimental results show that FastPR outperforms the state-of-the-art approaches in terms of both efficiency and accuracy, with over 15× inference time reduction and average 2% accuracy improvement.

Our code will be released in https://github.com/Sunmeng1997/FastPR.

2 RELATED WORK

This section briefly summarizes the related works in semantic person retrieval. Besides, we also introduce the existing studies in visual grounding that is similar with but has different focuses on semantic person retrieval, and recent advances in self-supervised learning that plays an important role in FastPR.

2.1 Semantic Person Retrieval

Existing solutions for semantic person retrieval mainly focus on using semantic descriptions to accurately find the specific person, either from cropped images with a single person [1, 17, 21, 31] or adopting a pretrained objected detector to output cropped images first [3, 23, 33]. Most of them aim to achieve higher accuracy of person retrieval by addressing the problem of cross-modal feature alignment. Sarafianos *et al.* [21] propose to use the adversarial loss to optimize the match between the image and textual features. Zheng *et al.* [32] design a new system to discriminatively embed the image and text to a shared visual-textual space. Niu *et al.* [17] declare the difficulty in directly measuring the similarity between images and descriptions due to the modality heterogeneity, and hence propose a Multi-granularity Image-text Alignments (MIA) model to alleviate the cross-modal fine-grained problem for better similarity evaluation. Recently, Zhou *et al.* [34] design a novel Deep Surroundings-person Separation Learning (DSSL) model by separating the information into surroundings and person to achieve a higher retrieval accuracy. While all the above works are based on cropped images with a single person, making them unable to be fully automatic in real applications, there are a few of works choose the full images as input. Zhou *et al.* [33] adopt a region proposal network in Faster R-CNN to generate the cropped images of person and extract the visual features, then integrate the visual and text features to score region proposals for generating the final output. Shah *et al.* [23] use Mask R-CNN to detect persons in an images and use a list of filters to locate the final target. All of the works adopt a pretrained objected detector to output cropped images first, which can be seen as two-stage frameworks. However, as we mentioned in Introduction, two-stage models suffer from heavy

time complexity and error accumulation in inference. Different from prior studies, FastPR is an end-to-end framework, which uses the orignial surveillance images as input and outputs the final localization of the target person by semantic descriptions directly.

2.2 Proxy Task in Self-supervised Learning

Self-supervised learning (SSL) has been wildly used in different contexts and fields, such as representation learning, natural language processing[4, 6], computer vision[5, 8, 18, 30] and reinforcement learning. Applying self-supervised learning can help the model have better generalization ability of the input. Generative self-supervised learning is one of the mainstream SSL methods [14]. The basic idea is to automatically generate some kinds of supervisory signals from the input and solve a specifically designed task (also called as proxy task). There are three kinds of proxy tasks:

(a) **Content-based proxy task** is mainly based on the context information of the data itself, to learn the characteristics of the data. This kind of proxy tasks is usually set to recover the input under some corruption, e.g., context encoder in computer vision [18], mask-based method like BERT [4]. Others tend to design some indirect tasks like predicting the "rotation" [8], patch order prediction [5], patch position prediction [12], colorization [30].

(b) **Temporal-based proxy task** is widely used in video-related applications. Sermanet *et al.* [22] claim that the adjacent frame features in the video are similar, while the video frames that are far apart are dissimilar, thus self-supervised constraints can be constructed by this. Misra *et al.* [15] design a self-supervised proxy task that determines whether a sequence of frames from a video is in the correct temporal order. Wu *et al.* [25] apply self-supervised learning in dialogue system by designing a proxy task to detect the dialogue flow.

(c) **Contrastive-based proxy task** is achieved by constructing positive and negative samples, and then measuring the distance between positive and negative samples. The core idea of this kind of proxy task is that the similarity between positive samples is far greater than the similarity between the negative samples. Tian *et al.* [24] learn a representation aiming to maximize mutual information between different views of the same scene but is otherwise compact, which employs a contrastive-based method for cross-modal tasks.

Different from the general proxy tasks mentioned above, we specifically design proxy tasks according to the challenges in real-world cross-modal person retrieval task. Since surveillance images with full background have much more information than the semantic descriptions, we set a semantic label prediction proxy task to guarantee the effect of fused features. To overcome the low resolution problem in real surveillance scenes, an upsampling reconstruction proxy task is designed to enhance the local feature of the target person. In addition, a novel offset prediction proxy task assists the localization network to have the ability of identify the target person in crowd.

3 METHOD

Given an image I and a list of semantic descriptions T, our goal is to localize the target people in I with T directly. Fig 2 shows the overview of FastPR. It consists of three key modules: **multimodal feature encoder**, **dynamic visual-semantic alignment**, and **dual-granularity person localization**. The input image and semantic descriptions are first fed into the feature encoders to extract the initial image and semantic features. Then, the dynamic visual-semantic alignment module fuses the cross-modal feature pairs, through the broadcast multiplication between the semantic and visual feature and softmax to get the initial fused features, a classification proxy task to constrain the effectiveness of fusion. The fused features are finally used by the dual-granularity person localization module to derive the target person's position. During the retrieval stage, we apply several specially designed self-supervised learning proxy tasks to assist the retrieval process. The design details of each module are presented as follows.

3.1 Multimodal Feature Encoders

An image encoder and a semantic encoder are designed to extract different modal features from the input I respectively. Different from the existing two-stage methods that can only extract person features from a cropped image, FastPR is designed to quickly extract all the person features from the original one. To achieve this, we divide the input image into $S * S$ grids, where a grid cell is regarded as containing the target person and responsible for person detection if the center of the person falls into this gird. We use a ResNet50 [10] pretrained on ImageNet[20] as the visual backbone network $\mathcal{R}$ to extract visual features $I_{en} \in \mathbb{R}^{M*S*S}$ from an input image, where $M = 256, S = 7$ in ResNet50. Formally, the encoder process is defined as:

$$I_{en} = \mathcal{R}(I).$$

For a list of semantic description phrases $E = \{e_1, e_2, ..., e_T\}$, we label each semantic description with a $d-$dimensional one-hot vector and turn them into semantic embedding v_i by a trainable embedding layer $\mathcal{E}$, i.e.,

$$v_i = \mathcal{E}(e_i).$$

Through the encoder part, we can obtain the initial representation of both image and semantic descriptions. Note that, we set the same dimension M for both visual and semantic features for the following grid-based alignment. The original input is turn into the visual-semantic feature pairs P, which is denoted as:

$$P = \{I_{en} \in \mathbb{R}^{M*S*S}, V = \{v_i \in \mathbb{R}^M\}_{i=0}^{T}\}.$$

3.2 Dynamic Visual-semantic Alignment

3.2.1 Attention-based Cross-modal Alignment. For a specific feature pair $p \in P$, we aim to fuse the visual-semantic feature appropriately and locate the target person by the fused feature. Given the unbalance of information between the visual and semantic features, we perform grid-based attention over them. As we regard the visual features as $S * S$ grids, for a specific semantic feature v_i, a weight map w_i will be generated as:

$$w_i = Softmax(I_{en} \otimes v_i),$$

where $\otimes$ denotes the matrix multiplication along M dimension, the $Softmax$ function maps the weights into $[0, 1]$. Elements with higher values in w_i indicates that the corresponding grid contains more information about the semantic feature v_i. For each $v_i \in V$ in a pair, we generate the fused visual-semantic features as:

$$\bar{I}_i = I_{en} \odot w_i,$$

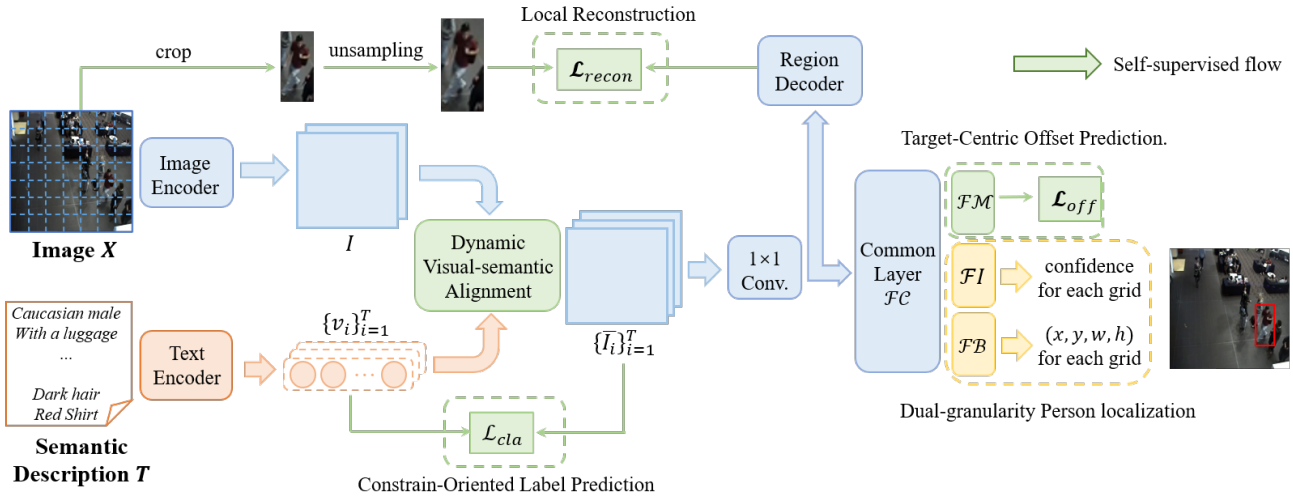

Figure 2: The overview of FastPR.

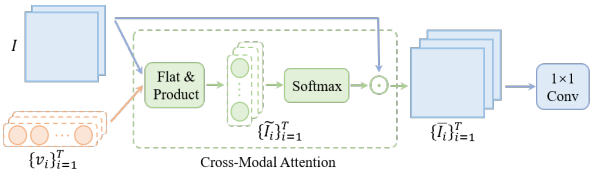

Figure 3: The overview of attention-based cross-modal alignment

where $\odot$ denotes element-wise multiplication and $\bar{I}_i$ denotes the visual feature I fusing with the semantic feature v_i. After the separate attention process, we perform a 1×1 convolutional layer C to integrate the separate features $\bar{I}_i \in \mathbb{R}^{S*S*M}$ as follows:

$$\bar{I_{en}} = (\bar{I}_1||\bar{I}_2||...||\bar{I}_T),$$
$$\bar{I} = C(\bar{I_{en}}),$$

where $||$ denotes the concatenation of $\bar{I}_i$ along the M dimension. The overview of cross-modal fusion module is shown in Fig. 3.

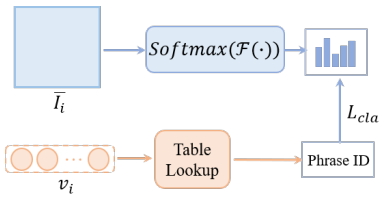

Figure 4: The overview of constrain-Oriented label prediction proxy task.

3.2.2 Constrain-Oriented Label Prediction. We design a prediction proxy task to ensure the effect of attention module and fill the gap between semantic features and visual features better. Specifically, for a fused feature $\bar{I}_i$, we feed it into a fully-connected layers $\mathcal{F}$ to predict the semantic feature v_i's corresponding one-hot vector in the dataset:

$$l_i = \mathcal{F}(\bar{I}_i),$$
$$\tilde{l}_i = Softmax(l_i),$$

where $\tilde{l}_i$ represents the prediction of the corresponding one-hot vector of v_i according to $\bar{I}_i$, which forces the fused feature to contain the information of the semantic phrase. We apply the categorical cross-entropy loss to optimize the label prediction process as below:

$$\mathcal{L}_{cla} = -\frac{1}{T}\sum_{t=0}^{T}(l_i ln\tilde{l_i} + (1 - l_i)ln(1 - \tilde{l_i})), \quad (1)$$

which l_i represents the ground truth one-hot vector of semantic phrase v_i. The fused feature is supposed to contain the corresponding semantic information by optimizing Equation (1). Fig 4 shows the overview of the constrain-oriented label prediction proxy task.

3.3 Dual-granularity Person Localization

3.3.1 Unsampling Local Reconstruction. After the cross-modal attention module, we can obtain the fused feature $\bar{I}$. However, since surveillance images often suffer from low resolution and blurriness, a reconstruction proxy task is designed to enhance the local features in background and noise. Specifically, we first extract the cropped image I_p of target person according to the ground truth of the bounding box GT and apply the bilinear interpolation algorithm $\mathcal{B}$ to unsample the local image of the target person, which can be denoted as:

$$I_p = Crop(I|GT),$$
$$\bar{I_p} = \mathcal{B}(I_p).$$

As shown in Fig 2, a trainable decoder $\mathcal{D}$ takes the fused feature $\bar{I}$ as input to generate a reconstruction of the enhanced local image $\bar{I_p}$, i.e.,

$$\tilde{I_p} = \mathcal{D}(\bar{I}).$$

To optimize the reconstruction process, we set up a reconstruction loss as:

$$\mathcal{L}_{rec} = \| \bar{I}_p - \tilde{I}_p \| . \tag{2}$$

The reconstruction proxy task enhances the fused feature to contain more information of the target, and hence promotes the fusion of cross-modal features indirectly.

3.3.2 Grid-level Dual-granularity Person localization. During the localization procedure, both the confidence prediction network and bounding box prediction network are supposed to filter the location information of the target from the input $\bar{I}$, thus, the shallow feature extraction should be similar. Based on that, we set a 2-layer fully connected layer $\mathcal{FC}$ as the common layer to extract the initial features.

We use the fused feature $\bar{I}$ to localize the target person. Different from the previous work on detection, we divide the localization in a dual-grained way. Specifically, we regard the prediction confidence of whether the target person falls into a grid as a coarse-grained prediction, while the bounding box prediction can be treated as a fine-grained localization. For the confidence prediction, a 2-layer fully connected layer $\mathcal{FI}$ after the $\mathcal{FC}$ is designed to generate the prediction of confidence C that can be denoted as:

$$C = \mathcal{FI}(\mathcal{FC}(\bar{I})).$$

Obviously, C has a size of $S * S$. For each grid, there is a ground truth $\hat{C}_i \in \{0, 1\}$ to denote whether this grid has the target or not. Then, the loss of confidence prediction is defined as:

$$\mathcal{L}_{con} = \sum_{i=0}^{S^2} (C_i - \hat{C}_i)^2. \tag{3}$$

For the prediction of bounding box, we feed the fused feature into another fully connected layer $\mathcal{FB}$ after the common layer $\mathcal{FC}$ and get the prediction of bounding box (x, y, w, h), where (x, y) are the coordinates representing the center of the box relative to the bounds of the grid cell, and the width w and height h are predicted relative to the whole image [19], respectively. The loss of the bounding box regression is defined as:

$$\begin{aligned} \mathcal{L}_{loc} = &\sum_{i=0}^{S^2} \mathbb{1}_i^{obj} ((x_i - \hat{x})^2 + (y_i - \hat{y})^2 + \\ &\sum_{i=0}^{S^2} \mathbb{1}_i^{obj} ((\sqrt{w_i} - \sqrt{\hat{w}})^2 + (\sqrt{h_i} - \sqrt{\hat{h}})^2)), \end{aligned} \tag{4}$$

where the $\hat{x}, \hat{y}$ are the ground truth of the coordinates representing the center of the box relative to the bounds of the grid cell, and the $\hat{w}, \hat{h}$ are the ground truth of the weight and height relative to the whole image, respectively. Besides, $\mathbb{1}_i^{obj}$ means that

$$\mathbb{1}_i^{obj} = \begin{cases} 1, & target \in grid_i \\ 0, & otherwise \end{cases}.$$

3.3.3 Target-Centric Offset Prediction. In order to make the localization network capable of identifying and distinguishing the target person in the crowd, we propose a specially designed offset proxy task for the common layer $\mathcal{FC}$ to enhance the layer's perception of the target person location.

Specifically, we divide the $S * S$ grids into 9 regions. As shown in Fig 5, for each grid $g_i, i = \{1, 2, ..., S * S\}$ if it contains the target person in the region $r_q, q = \{0, 2, ..., 8\}$, we randomly swap the g_i's feature with the grid in other regions, e.g., r_p. The offset is calculated as:

$$o_{qp} = |q - p|.$$

where $|\cdot|$ denotes the absolute value. There are eight different offset values in this setting, which means this proxy task is a multi-class classification problem.

For the fused feature $\bar{I}$ with the dimension of $\mathbb{R}^{S*S*M}$, we swap the grid features corresponding to the target person with the features in other grids and get the swapped feature $\bar{I}_s$. A 1-layer MLP network $\mathcal{FM}$ is designed to predict the absolute value of the offset between exchanged and original features after the common layer $\mathcal{FC}$, which is denoted as:

$$\tilde{o} = \mathcal{FM}(\mathcal{FC}(\bar{I}_s)).$$

Since the location network is supposed to identify the target person in spite of his location exchanging in the image, the offset between the $\bar{I}$ and $\bar{I}_s$ should be predicted precisely. The loss function of the offset prediction proxy task is defined as:

$$\mathcal{L}_{off} = -(olog\tilde{o} + (1 - o)log(1 - \tilde{o})), \tag{5}$$

where o is the ground truth of the offset.

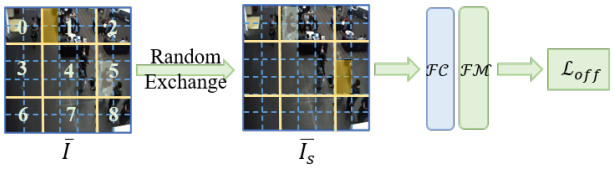

Figure 5: The overview of target-centric offset prediction proxy task. The fused feature is divided into 9 regions, the grid feature where the target is located is arbitrarily exchanged with other grid features, and the offset is the difference between the region numbers to which the two sets of features belong. E.g., the first swapped feature's absolute value of the offset is $|5 - 4| = 1$, and the second is $|5 - 6| = 1$. There are 8 different offsets in this setting.

3.4 Optimization and Inference

3.4.1 Optimization for training. The total loss for training FastPR can be described as:

$$L = \lambda_{ssl} L_{ssl} + \lambda_{con} L_{con} + \lambda_{loc} L_{loc}, \tag{6}$$

where we use L_{ssl} to denote the total loss of all the self-supervised proxy tasks, which contains three self-supervised loss terms and can be calculated as follows:

$$L_{ssl} = \lambda_{cla} L_{cla} + \lambda_{recon} L_{recon} + \lambda_{off} L_{off}, \tag{7}$$

where $\lambda_{ssl}, \lambda_{con}, \lambda_{loc}, \lambda_{cla}, \lambda_{recon}$ and λ_{off} denote the hyperparameters to balance different loss terms during training.

As the proxy tasks are designed to achieve more accurate semantic person retrieval performance, they may constrain the model in different aspects. Thus, we choose to optimize the loss of the localization and proxy tasks alternately. Specifically, we start the

Table 1: Performance comparison of FastPR to the related methods training and testing on SoftBioSearch Dataset. Experiment details can be seen in Section 4.3

Model	R@1 (%)			R@5 (%)			Time(ms)
	IoU >0.4	*IoU >0.5*	*IoU >0.6*	*IoU >0.4*	*IoU >0.5*	*IoU >0.6*	
Cos-Sim	33.50	29.08	25.56	34.89	32.46	26.69	280
Attn-Based [33]	49.98	43.80	40.85	58.85	54.76	50.04	355
Trans-SPR [27]	53.90	51.10	44.68	57.46	54.57	50.94	295
Per-Vis [23]	**67.80**	57.30	55.79	70.25	62.18	57.80	320
FastPR	66.02	**60.14**	**56.82**	**71.73**	**64.07**	**58.30**	**17**

optimization with proxy tasks for several epochs and freeze the parameters of the confidence prediction network and bounding box prediction network, then exchange the frozen order for another several eopchs.

3.4.2 Inference. During the inference, both of the confidence prediction network and bounding box prediction network will generate the prediction of confidence and bounding box in each grid according to the semantic descriptions. The max of predicted confidence decides which grid is responsible for the target, and the corrosponding predicted bounding box of this grid is seen as the finally output.

4 IMPLEMENTATION AND EVALUATION

4.1 Training and Evaluation Metrics

Training Details. In our experiments, we pretrain the RestNet50 [10] on ImageNet as backbone to get image representations. To this end, each image is resized into the size of $224 * 224$. For semantic descriptions, text embeddings are regarded as trainable parameters. We use the Adam [11] optimizer and our set the learning rate into 1e-3. We train our model using the batch size of 256 on an A40 Nvidia Graphic card for 300 epochs. In our loss function, the temperature parameter τ is set as 0.5, and the $\lambda_{ssl}, \lambda_{con}, \lambda_{loc}$ are set as $1, 100, 5$ respectively. During the inference process, an image and a list of semantic description are fed into the network and the confidence prediction network will generate the confidence of each grid. We choose the grid which has the greatest confidence, and pick this grid's bounding box as the final output if the greatest confidence are larger than 0.6, otherwise we think the image doesn't contain the target person.

Evaluation Metrics. We adopt the recall $R@n, IoU > m$ as the evaluation metric, which is widely used in retrieval tasks. Recall R@n is calculated by the the percentage of samples for which the correct result resides in the top-n retrievals to the query. The result is seen as correct when the IoU between the prediction bounding box and the ground truth is greater than m. The higher value of m donates the stricter criteria of retrieval. We set the $n = 1, 5, m = 0.4, 0.5, 0.6$ to see the different results in experiment. For the evaluation of efficiency, we record the computation time of a batch (set with 8) while performing inference on an A40 GPU. We run different model with 10 times and get the average of the ten running times as the final result.

4.2 Dataset and Baselines

Dataset. We use the SoftBioSearch [9] dataset as our training and testing set. The dataset are collected from 6 stationary cameras. Each sequence contains detailed semantic information for a single search subject who appears in the clip (gender, age, height, build, hair and skin colour, clothing type, texture and colour), and are annotated with the target subject location (over 11, 000 frames are annotated in total). The training set and test set were randomly divided according to the ratio of 7 to 1.

Baselines. We compare the performance of FastPR with state-of-the-art semantic person retrieval methods using the surveillance images of the aforementioned dataset. The four methods used in perfromance comparison are introduced as follows.

- *Cos-Sim*, which is a simple two-stage method. It adopts a Faster R-CNN object detector for person detection and directly compute cosine similarities between salient proposals and semantic label representations. The bounding box with highest similarity is retrieved as the final result.
- *Attn-Based*, indicating the attention-based model proposed by Zhou et al. [33]. Compared with *Cos-Sim*, they also use Faster R-CNN object detector for person detection. However, the cross-modal attention mechanism is applied for region selection. Since their dataset and implementations are not released to public, we reproduce and train a model under same settings based on our dataset for fair comparison.
- *Trans-SPR* [27] , which introduces transfer learning to find more features for region selection. During the person detection process under Mask R-CNN, they leverage a pre-trained DenseNet-161 network to predict multiple aspects of attributes (e.g., gender, pose, luggage, clothings) under transfer learning techniques. These features are further used for computing the matching score between proposals and semantic labels for person retrieval. Since the code is not open-sourced, we re-implement them based on the model descriptions of the original papers.
- *PeR-Vis* [23], which uses a cascade filtering of person descriptors to narrow down the search space. More specifically, it performs person detection under Mask R-CNN. During the retrieval process, cascade filters based on Height, Torso Module, Leg Module and Gender are adopted for more detailed and distinguishable selecting. This model reached previous SOTA results and are regarded as our strong baseline model.

4.3 Quantitative Performance

Table 1 shows the performance comparison results between FastPR and the baselines.

Table 2: Ablation study trained and tested on SoftBioSearch Dataset.

L_{cla}	L_{recon}	L_{off}	L_{con}	R@1 (%)			R@5 (%)		
				IoU >0.4	IoU >0.5	IoU >0.6	IoU >0.4	IoU >0.5	IoU >0.6
				51.70	47.26	40.79	55.98	50.46	43.26
✓				52.90	48.07	41.85	57.52	52.68	46.29
	✓			51.98	47.72	41.00	56.98	51.84	45.71
		✓		52.49	47.95	41.59	57.25	52.39	45.91
			✓	49.36	45.27	40.87	52.53	49.55	41.88
	✓	✓	✓	63.56	57.95	46.28	68.42	61.96	56.01
✓		✓	✓	64.58	58.69	54.79	69.28	62.13	56.25
✓	✓		✓	64.67	58.74	54.34	69.45	62.72	56.28
✓	✓	✓		60.58	53.16	49.77	63.79	55.59	51.89
✓	✓	✓	✓	**66.02**	**60.14**	**56.82**	**71.73**	**64.07**	**58.30**

4.3.1 Efficiency performance. Other methods run in around 300ms in average, while FastRP can achieve 17ms. The two-stage method spends much time on detecting persons while most of them aren't the target, costing most of the inference time. FastPR accomplish the retrieve task in integrating one-stage way, making it much more efficient than the previous methods.

4.3.2 Accuracy Performance. First, we compare our model with all baseline models, especially the strong baseline Per-Vis. For *R*@5, our model shows significant improvement under each metric. For *R*@1, FastPR gets comparable results compared with Per-Vis under $IoU > 0.4$. It significantly outperforms Per-Vis on $IoU > 0.5, IoU > 0.6$ which are more difficult tasks. This illustrates that our model gets more precise bounding boxes compared with prior works.

Among baseline models, there is a gap between *Cos-Sim* with others, since the retrieval process is simply based on cosine distances. This does not take the natural gap between visual and semantic features into account. Using attention mechanism for dynamic soft selection, *Attn-Based* model get much better performances. Compare with the two baselines, *Trans-SPR* introduces more features for retrieval, while *Per-Vis* focuses on using cascade filtering to narrow down the search space, i.e., performing deeper cross-modal interaction. Both of the two models get competable results, and *Per-Vis* shows more improvements than *Trans-SPR*. It turns out that compared with finding more features, exploring better retrieving method, i.e., performing deeper cross-modal interaction, has more significant improvement. To this end, we leverage the idea of self-supervised learning and introduce additional proxy tasks for deeper cross-modal interaction. Compared with *Trans-SPR* using transfered features and *Per-Vis* leveraging carefully-designed cascade filtering (e.g., cloth color, hair color, etc.), our self-supervised method does not use any human-annotated data and requires less hand-crafted filter engineering.

4.4 Case Study

4.4.1 Accurate Retrieval. The qualitative performance of FastPR is shown in Fig. 6. The first row shows that FastPR could locate the target accurately in a real-world environment. Besides, according to the second example in the third row, FastPR is capable of localizing the person accurately when the target is partially occluded or missing. In addition, in the second row which the images contain a large number of people, FastPR can filter out the incompatible objects according to the semantic descriptions, and finally obtain accurate coordinates. It demonstrates that FastPR can also localize the target task accurately in the presence of complex backgrounds and crowds.

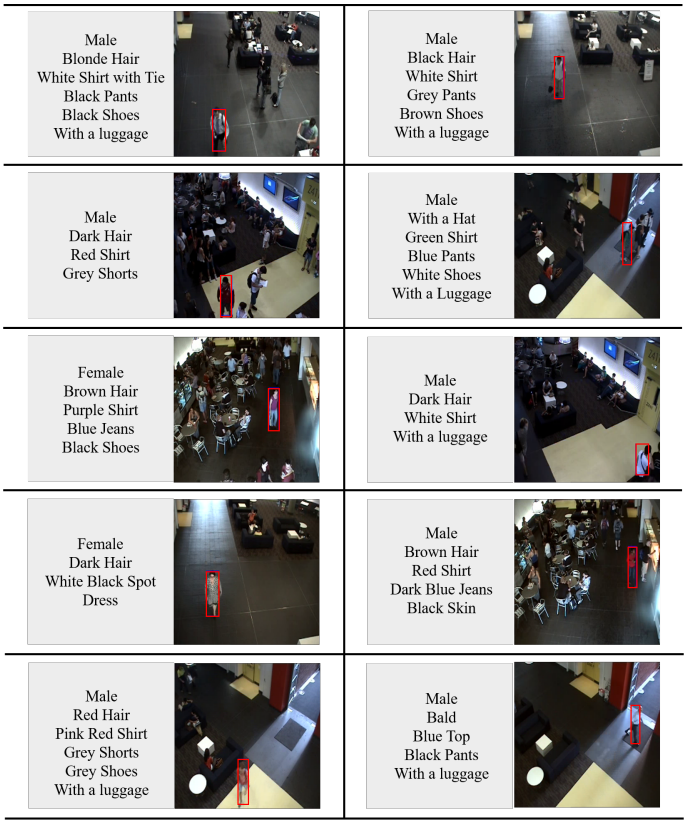

Figure 6: Case study of FastPR.

4.4.2 Fuzzy Retrieval. We also evaluate the performance of FastPR on fuzzy retrieval in Fig. 7. When the semantic descriptions are incomplete to retrieve a certain person, FastPR can detect all the

Figure 7: The qualitative performance of FastPR with fuzzy retrieval.

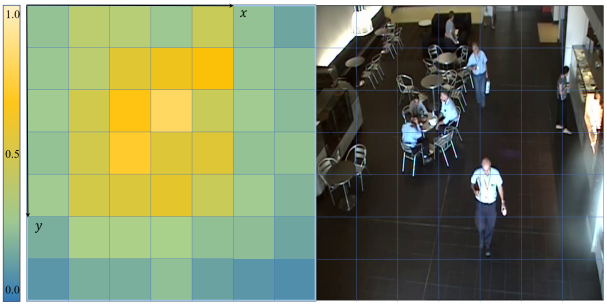

Figure 8: The qualitative performance of confidence. Set "Blue Shirt" as textual description.

eligible candidates. We use different semantic descriptions including clothing, gender and hair color, etc., FastPR can exhibit stable retrieval accuracy.

4.4.3 Confidence Performance. We also show the output of the confidence prediction layer in Fig. 8. Specifically, we only set "Blue Shirt" as the semantic description input, and the confidence prediction layer $\mathcal{FC}$ output the probability C_i of each grid that contains a person matching the description of "Blue Shirt". We can clearly see from the Fig. 8 that in the heat map of confidence, the grids with a person in "Blue Shirt" have significantly higher scores. It demonstrates that the confidence predicting layer $\mathcal{FC}$ could understand and perceive the emphasis of semantic information on pictures.

4.5 Ablation Studies

To achieve more accurate semantic person retrieval, we design three self-supervised learning proxy tasks for different aspects of the problem. Besides the self-supervised learning part, we also divide the regression of localization into dual-granularity detection, including confidence prediction and bounding box prediction. In the ablation study, We delete the three proxy tasks and train the FastPR separately to see the effectiveness of the different proxy tasks. Besides, we merge the confidence and bounding box prediction into one localization network to replace the original dual-granularity person localization. The results of the ablation study are in Table 2.

4.5.1 The effectiveness of self-supervised proxy tasks. There are three different proxy tasks in our model. The classification task is to supervise the cross-modal attention fusion module, and the reconstruction task force the fused feature to contain more information of the target person, the offset prediction task enhance the ability of detection in the localization network. We can see the different results of whether use these proxy tasks or not. FastPR can only reach around 47% by using one proxy task, and have around +6% improvement by using three proxy tasks, indicating the effectiveness of the specific-designed self-supervised proxy tasks.

4.5.2 The effectiveness of confidence prediction network. In the localization task, due to the complexity of semantic retrieval task, the confidence and exact bounding box are not predicted by the same fully connected network as YOLO [19]. Table 2 shows that $R@1$ and $R@5$ both experience a significant decline if we predict the confidence and bounding box in the same network, which proves the effectiveness of the separate design.

5 CONCLUSION

In this work, we propose a one-stage semantic person retrieval method, FastPR, to achieve fast and accurate person retrieval from real-world surveillance images. We design a dynamic visual-semantic alignment mechanism that to fuse the cross-modal features, and a label prediction proxy task to constrain the attention process. Real-world surveillance images usually suffer from low resolution problem, an unsampling local reconstruction task is designed to enhance the local target feature in the fused feature. We propose dual-granularity person localization module to precisely detect the target person by semantic descriptions. To make the localization network capable of identifying and distinguishing the target person in the crowd, we propose a specially designed offset proxy task. Experimental results on a surveillance image dataset SoftBioSearch[9] demonstrate that FastPR outperforms the state-of-the-art semantic person retrieval methods in terms of both efficiency and retrieval accuracy.

6 ACKNOWLEDGEMENT

This research was supported in part by the National Natural Science Foundation of China under Grant No. 62122095, 62072472 and U19A2067, by a grant from the Guoqiang Institute, Tsinghua University, and the Natural Science Foundation of Hunan Province, China under Grant No. 2020JJ2050.

REFERENCES

[1] Surbhi Aggarwal, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. 2020. Text-based person search via attribute-aided matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2617–2625.

[2] Donghyeon Baek, Youngmin Oh, and Bumsub Ham. 2021. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9536–9545.

[3] Simon Denman, Michael Halstead, Clinton Fookes, and Sridha Sridharan. 2015. Searching for people using semantic soft biometric descriptions. *Pattern Recognition Letters* 68 (2015), 306–315.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*. 1422–1430.

[6] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. 2018. Weakly supervised dense event captioning in videos. *Advances in Neural Information Processing Systems* 31 (2018).

[7] Benjamin Elizalde, Shuayb Zarar, and Bhiksha Raj. 2019. Cross modal audio search and retrieval with joint embeddings based on text and audio. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4095–4099.

[8] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728* (2018).

[9] Michael Halstead, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2014. Locating people in video from semantic descriptions: A new database and approach. In *2014 22nd International Conference on Pattern Recognition*. IEEE, 4501–4506.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[11] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[12] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. 2019. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1920–1929.

[13] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. 2019. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1950–1959.

[14] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2021. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[15] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. 2016. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*. Springer, 527–544.

[16] Niluthpol Chowdhury Mithun, Rameswar Panda, Evangelos E Papalexakis, and Amit K Roy-Chowdhury. 2018. Webly supervised joint embedding for cross-modal image-text retrieval. In *Proceedings of the 26th ACM international conference on Multimedia*. 1856–1864.

[17] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. 2020. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing* 29 (2020), 5542–5556.

[18] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2536–2544.

[19] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

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

[21] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. 2019. Adversarial representation learning for text-to-image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5814–5824.

[22] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. 2018. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 1134–1141.

[23] Parshwa Shah, Arpit Garg, and Vandit Gajjar. 2021. Per-vis: Person retrieval in video surveillance using semantic description. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 41–50.

[24] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *European conference on computer vision*. Springer, 776–794.

[25] Jiawei Wu, Xin Wang, and William Yang Wang. 2019. Self-supervised dialogue learning. *arXiv preprint arXiv:1907.00448* (2019).

[26] Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. 2020. Cross-modal attention with semantic consistence for image–text matching. *IEEE transactions on neural networks and learning systems* 31, 12 (2020), 5412–5425.

[27] Takuya Yaguchi and Mark S. Nixon. 2018. Transfer Learning Based Approach for Semantic Person Retrieval. In *15th IEEE International Conference on Advanced Video and Signal-based Surveillance: AVSS 2018 (30/11/18)*.

[28] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10502–10511.

[29] Tan Yu, Yi Yang, Yi Li, Lin Liu, Hongliang Fei, and Ping Li. 2021. Heterogeneous attention network for effective and efficient cross-modal retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1146–1156.

[30] Richard Zhang, Phillip Isola, and Alexei A Efros. 2017. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1058–1067.

[31] Ying Zhang and Huchuan Lu. 2018. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*. 686–701.

[32] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 2 (2020), 1–23.

[33] Tao Zhou, Muhao Chen, Jie Yu, and Demetri Terzopoulos. 2017. Attention-based natural language person retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 27–34.

[34] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. 2021. DSSL: Deep Surroundings-person Separation Learning for Text-based Person Retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*. 209–217.