Gallery Filter Network for Person Search

Lucas Jaffe
UC Berkeley
lucasjaffe@berkeley.edu

Avideh Zakhor
UC Berkeley
avz@berkeley.edu

Abstract

In person search, we aim to localize a query person from one scene in other gallery scenes. The cost of this search operation is dependent on the number of gallery scenes, making it beneficial to reduce the pool of likely scenes. We describe and demonstrate the Gallery Filter Network (GFN), a novel module which can efficiently discard gallery scenes from the search process, and benefit scoring for persons detected in remaining scenes. We show that the GFN is robust under a range of different conditions by testing on different retrieval sets, including cross-camera, occluded, and low-resolution scenarios. In addition, we develop the base SeqNeXt person search model, which improves and simplifies the original SeqNet model. We show that the SeqNeXt+GFN combination yields significant performance gains over other state-of-the-art methods on the standard PRW and CUHK-SYSU person search datasets. To aid experimentation for this and other models, we provide standardized tooling for the data processing and evaluation pipeline typically used for person search research.

1. Introduction

In the *person search* problem, a *query* person image crop is used to localize co-occurrences in a set of scene images, known as a *gallery*. The problem may be split into two parts: 1) *person detection*, in which all person bounding boxes are localized within each gallery scene and 2) *person re-identification* (re-id), in which detected gallery person crops are compared against a query person crop. Two-step person search methods [5, 9, 13, 19, 32, 40] tackle each of these parts explicitly with separate models. In contrast, end-to-end person search methods [2–4, 6, 8, 12, 14, 18, 20, 21, 23, 26, 34–39, 41] use a single model, typically sharing backbone features for detection and re-identification.

For both model types, the same steps are needed: 1) computation of detector backbone features, 2) detection of person bounding boxes, and 3) computation of feature embeddings for each bounding box, to be used for retrieval. Improvement of person search model efficiency is typically

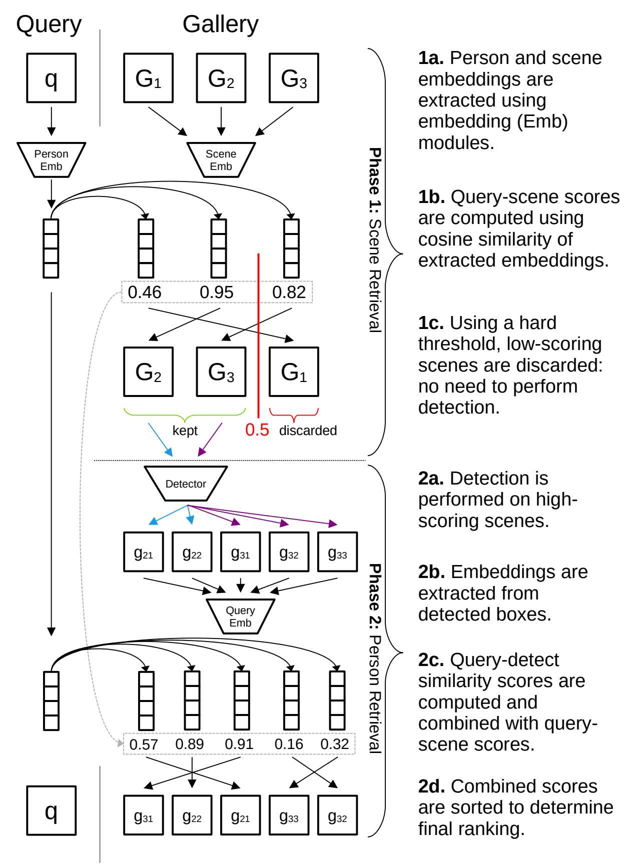

Figure 1: An illustration of our proposed two-phase retrieval inference pipeline. In the first phase, the Gallery Filter Network discards scenes unlikely to contain the query person. The second phase is the standard person retrieval process, in which persons are detected, corresponding embeddings extracted, and these embeddings are compared to the query to produce a ranking.

focused on reducing the cost of one or more of these steps. We propose the second and third steps can be avoided altogether for some subset of gallery scenes by splitting the retrieval process into two phases: scene retrieval, followed by typical person retrieval. This two-phase process is visualized in Figure 1. We call the module implementing scene retrieval the Gallery Filter Network (GFN), since its function is to filter scenes from the gallery.

By performing the cheaper query-scene comparison before detection is needed, the GFN allows for a modular computational pipeline for practical systems, in which one process can determine which scenes are of interest, and another can detect and extract person embeddings only for interesting scenes. This could serve as an efficient filter for video frames in a high frame rate context, or to cheaply reduce the search space when querying large image databases.

The GFN also provides a mechanism to incorporate global context into the gallery ranking process. Instead of combining global context features with intermediate model features as in [9, 20], we explicitly compare global scene embeddings to query embeddings. The resulting score can be used not only to filter out gallery scenes using a hard threshold, but also to weight predicted box scores for remaining scenes.

We show that both the hard-thresholding and score-weighting mechanisms are effective for the benchmark PRW and CUHK-SYSU datasets, resulting in state-of-the-art retrieval performance (+2.7% top-1 accuracy on the PRW dataset over previous best model), with improved efficiency (over 50% per-query cost savings on the CUHK-SYSU dataset vs. same model without the GFN). Additionally, we make contributions to the data processing and evaluation frameworks that are used by most person search methods with publicly available code. That work is described in Supplementary Material Section A.

1.1. Contributions

Our contributions are as follows:

- The Gallery Filter Network: A novel module for learning query-scene similarity scores which efficiently reduces retrieval gallery size via hard-thresholding, while improving detected embedding ranking with global scene information via score-weighting.
- Performance improvements and removal of unneeded elements in the SeqNet person search model [21], dubbed SeqNeXt.
- Standardized tooling for the data pipeline and evaluation frameworks typically used for the PRW and CUHK-SYSU datasets, which is extensible to new datasets.

All of our code and model configurations are made publicly available[1].

2. Related Work

Person Search. Beginning with the release of two benchmark person search datasets, PRW [40] and CUHK-SYSU [35], there has been continual development of new deep learning models for person search. Most methods utilize the Online Instance Matching (OIM) Loss from [35] for the re-id feature learning objective. Several methods [20,36,39] enhance this objective using variations of a triplet loss [29].

Many methods make modifications to the object detection sub-module. In [2, 20, 36], a variation of the Feature Pyramid Network (FPN) [22] is used to produce multi-scale feature maps for detection and re-id. Models in [2, 36] are based on the Fully-Convolutional One-Stage (FCOS) detector [30]. In COAT [38], a Cascade R-CNN-style [1] transformer-augmented [31] detector is used to refine box predictions. We use a variation of the single-scale two-stage Faster R-CNN [28] approach from the SeqNet model [21].

Query-Based Search Space Reduction. In [3, 23], query information is used to iteratively refine the search space within a gallery scene until the query person is localized. In [9], Region Proposal Network (RPN) proposals are filtered by similarity to the query, reducing the number of proposals for expensive RoI-Pooled feature computations. Our method uses query features to perform a coarser-grained but more efficient search space reduction by filtering out full scenes before expensive detector features are computed.

Query-Scene Prediction. In the Instance Guided Proposal Network (IGPN) [9], a global relation branch is used for binary prediction of query presence in a scene image. This is similar in principal to the GFN prediction, but it is done using expensive intermediate query-scene features, in contrast to our cheaper modular approach to the task.

Backbone Variation. While the original ResNet50 [16] backbone used in SeqNet and most other person search models has been effective to date, many newer architectures have since been introduced. With the recent advent of vision transformers (ViT) [10] and a cascade of improvements including the Swin Transformer [24] and the Pyramid Vision Transformer (v2) [33], used by the PSTR person search model [2], transformer-based feature extraction has increased in popularity. However, there is still an efficiency gap with CNN models, and newer CNNs including ConvNeXt [25] have closed the performance gap with ViT-based models, while retaining the inherent efficiency of convolutional layers. For this reason, we explore ConvNeXt for our model backbone as an improvement to ResNet50, which is more efficient than ViT alternatives.

3. Methods

3.1. Base Model

Our base person search model is an end-to-end architecture based on SeqNet [21]. We make modifications to the model backbone, simplify the two-stage detection pipeline, and improve the training recipe, resulting in superior performance. Since the model inherits heavily from SeqNet, and uses a ConvNeXt base, we refer to it simply as SeqNeXt to distinguish it from the original model. Our model, com-

[1]Project repository: https://github.com/LukeJaffe/GFN

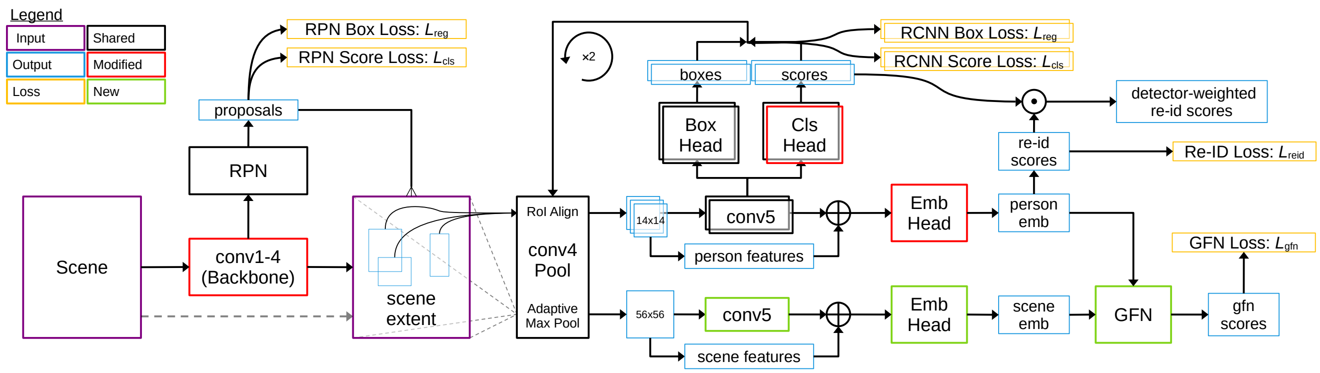

Figure 2: Architecture of the SeqNeXt person search model augmented with the GFN. Modules modified from SeqNet are colored red, and new modules, related to the GFN, are colored green. The model follows the standard Faster R-CNN paradigm, with backbone features from `conv4` being used to generate proposals via the RPN. `conv4` features are pooled for RPN proposals and passed through the `conv5` head to generate refined proposals. This process is repeated with the refined proposals to generate the final boxes. `conv4` features are also used to generate both person embeddings and scene embeddings in the same way: the person box or scene passes through the pooling block and then a duplicated `conv5` head, and `conv4`, `conv5` features are concatenated and passed through an embedding (Emb) head. In the pooling block, RoI Align [15] is used for person and proposal features, while adaptive max pooling is used for scene features. GFN scores are generated using person and scene embeddings from the same or different scenes. Person re-id scores are combined with the score output of the second R-CNN stage to produce detector-weighted scores.

bined with the GFN module, is shown in Figure 2.

Backbone Features. Following SeqNet's usage of the first four CNN blocks (`conv1-4`) from ResNet50 for backbone features, we use the analogous layers in terms of downsampling from ConvNeXt, also referred to as `conv1-4` for convenience.

Multi-Stage Refinement and Inference. We simplify the detection pipeline of SeqNet by duplicating the Faster R-CNN head [28] in place of the Norm-Aware Embedding (NAE) head from [6]. We still weight person similarity scores using the output of the detector, but use the second-stage class score instead of the first-stage as in SeqNet. This is depicted in Figure 2 as "detector-weighted re-id scores".

Additionally during inference, we do not use the Context Bipartite Graph Matching (CBGM) algorithm from SeqNet, discussed in Supplementary Material Section E.

Augmentation. Following resizing images to 900×1500 (Window Resize) at training time, we employ one of two random cropping methods with equal probability: 1) Random Focused Crop (RFC): randomly take a 512×512 crop in the original image resolution which contains at least one known person, 2) Random Safe Crop (RSC): randomly crop the image such that all persons are contained, then resize to 512×512. This cropping strategy allowed us to train with larger batch sizes, while benefiting performance with improved regularization. At inference time, we resize to 900×1500, as in other models. We also consider a variant of Random Focused Crop (RFC2), which resizes images so the "focused" person box is not clipped.

Objective. As in other person search models, we employ the Online Instance Matching (OIM) Loss [35], represented as $\mathcal{L}_{\text{reid}}$. This is visualized in Figure 3a. For all diagrams in Figure 3, we borrow from the spring analogy for metric learning used in *DrLIM* [11], with the concept of *attractions* and *repulsions*.

The detector loss is the sum of classification and box regression losses from the RPN, and the two Faster R-CNN stages, expressed as:

$$\mathcal{L}_{\text{det}} = \sum_{m \in M} \mathcal{L}^m_{\text{cls}} + \mathcal{L}^m_{\text{reg}}, \quad M = \{\text{RPN}, \text{RCNN1}, \text{RCNN2}\} \quad (1)$$

The full loss is the sum of the detector, re-id, and GFN losses:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{reid}} + \mathcal{L}_{\text{gfn}} \quad (2)$$

3.2. Gallery Filter Network

Our goal is to design a module which removes low-scoring scenes, and reweights boxes from higher-scoring scenes. Let s_{reid} be the cosine similarity of a predicted gallery box embedding with the query embedding, s_{det} be the detector box score, s_{gfn} be the cosine similarity for the corresponding gallery scene from the GFN, $\sigma(x) = \frac{e^{-x}}{1+e^{-x}}$, α be a temperature constant, and λ_{gfn} be the GFN score threshold. At inference time, scenes scoring below λ_{gfn} are removed, and detection is performed for remaining scenes, with the final score for detected boxes given by $s_{\text{final}} = s_{\text{reid}} \cdot s_{\text{det}} \cdot \sigma(s_{\text{gfn}}/\alpha)$.

The module should discriminate as many scenes below λ_{gfn} as possible, while positively impacting the scores of boxes from any remaining scenes. To this end, we consider

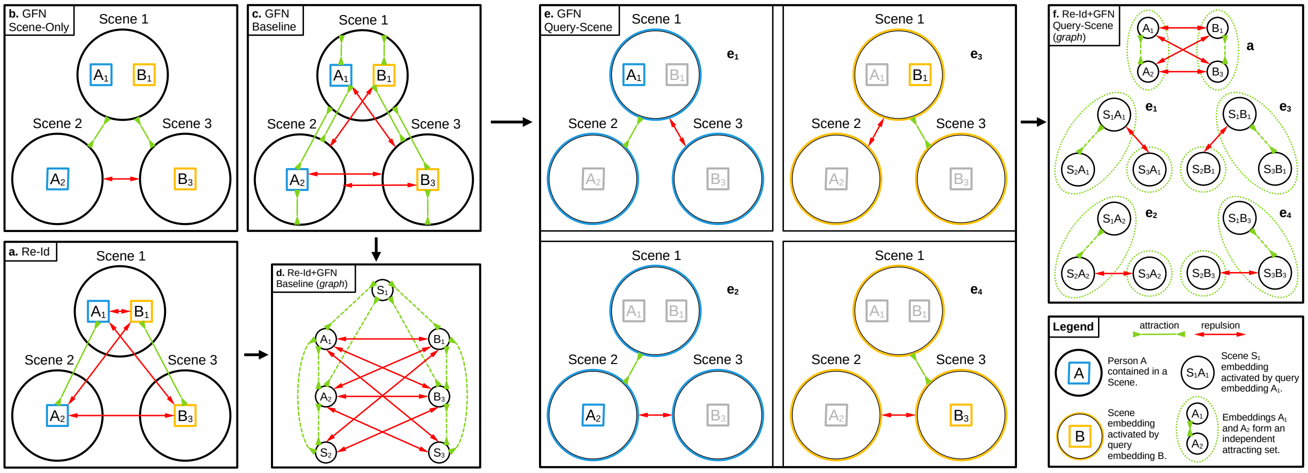

Figure 3: Visual representation of the re-id and GFN optimization objectives. In a), b), c), e), circles represent scene images which contain one or more different person identities, labeled A and B. We show a system of three scenes with two unique person identities. Green connectors represent attraction, meaning two embeddings are pushed together by an objective, and red connectors represent repulsion, meaning two embeddings are pulled apart by an objective. In a) we show the standard re-id loss objective. In b) we show the scene-only GFN objective. In c) we show the baseline GFN objective, and in e) we show the combined query-scene GFN objective. In d) we show the graph form of the baseline GFN objective and re-id objective together, and in f) we show the graph form of the combined query-scene GFN objective and re-id objective together, with green ellipses surrounding independent sets in each multipartite component.

three variations of the standard contrastive objective [7, 27] in Sections 3.2.1-3.2.3, in addition to a number of architectural and optimization considerations in Section 3.2.4.

3.2.1 Baseline Objective

The goal of the baseline GFN optimization is to push person embeddings toward scene embeddings when a person is contained within a scene, and to pull them apart when the person is not in the scene, shown in Figure 3c.

Let $x_i \in \mathbb{R}^d$ denote the embedding extracted from person q_i located in some scene s_j. Let $y_j \in \mathbb{R}^d$ denote the embedding extracted from scene s_j. Let X be the set of all person embeddings x_i, and Y the set of all scene embeddings y_j, with $N = |X|, M = |Y|$.

We define the query-scene indicator function to denote positive query-scene pairs as

$$\mathcal{I}_{i,j}^{Q} = \begin{cases} 1 & \text{if } q_i \text{ present in } s_j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

We then define a set to denote indices for a specific positive pair and all negative pairs: $K_{i,j}^{Q} = \{k \in 1, \ldots, M \,|\, k = j \text{ or } \mathcal{I}_{i,j}^{Q} = 0\}$. Define $\text{sim}(u, v) = u^\top v / \|u\| \|v\|$, the cosine similarity between two $u, v \in \mathbb{R}^d$, and τ is a temperature constant. Then the loss for a positive query-scene pair is the cross-entropy loss

$$\ell_{i,j}^{Q} = -\log \frac{\exp\left(\text{sim}(x_i, y_j)/\tau\right)}{\sum_{k \in K_{i,j}^{Q}} \exp\left(\text{sim}(x_i, y_k)/\tau\right)} \tag{4}$$

The baseline Gallery Filter Network loss sums positive pair losses over all query-scene pairs:

$$\mathcal{L}_{\text{gfn}}^{Q} = \sum_{i=1}^{N} \sum_{j=1}^{M} \mathcal{I}_{i,j}^{Q} \ell_{i,j}^{Q} \tag{5}$$

3.2.2 Combined Query-Scene Objective

While it is possible to train the GFN directly with person and scene embeddings using the loss in Equation 5, we show that this objective is ill-posed without modification. The problem is that we have constructed a system of opposing attractions and repulsions. We can formalize this concept by interpreting the system as a graph $G(V, E)$, visualized in Figure 3d. Let the vertices V correspond to person, scene, and/or combined person-scene embeddings, where an edge in E (red arrow) connecting any two nodes in V represents a negative pair used in the optimization objective. Let any group of nodes connected by green dashed arrows (not edges in G) be an independent set, representing positive pairs in the optimization objective. Then, each connected component of G must be multipartite, or the optimization problem will be ill-posed by design, as in the baseline objective.

To learn whether a person is contained within a scene while preventing this conflict of attractions and repulsions, we need to apply some unique transformation to query and scene embeddings before the optimization. One such option is to combine a query person embedding separately with the query scene and gallery scene embeddings to produce fused representations. This allows us to disentangle the web of interactions between query and scene embeddings, while still learning the desired relationship, visualized in Figure 3e. The person embedding used to fuse with each scene embedding in a pair is left colored, and the corresponding scenes are colored according to that person embedding. Person embeddings present in scenes which are not used are grayed out.

In the graph-based presentation, shown in Figure 3f, this modified scheme using query-scene embeddings will always result in a graph comprising some number of star graph connected components. Since these star graph components are multipartite by design, the issue of conflicting attractions and repulsions is avoided.

To combine a query and scene embedding into a single query-scene embedding, we define a function $f : \mathbb{R}^d, \mathbb{R}^d \rightarrow \mathbb{R}^d$, such that $z_{i,j} = f(x_i, y_j)$ and $w_i = f(x_i, y^{x_i})$, where y^{x_i} is the embedding of the scene that person i is present in. Borrowing from SENet [17] and QEEPS [26], we choose a sigmoid-activated elementwise excitation, with $\odot$ used for elementwise product. "BN" is a Batch Normalization layer, to mirror the architecture of the other embedding heads, and β is a temperature constant.

$$f(x, y) = \text{BN}(\sigma(x/\beta) \odot y) \tag{6}$$

Other choices are possible for f, but the elementwise-product is critical, because it excites the features most relevant to a given query within a scene, eliciting the relationship shown in Figure 3e.

The loss for a positive query-scene pair is the cross-entropy loss

$$\ell^C_{i,j} = -\log \frac{\exp\left(\text{sim}(w_i, z_{i,j})/\tau\right)}{\sum_{k \in K^Q_{i,j}} \exp\left(\text{sim}(w_i, z_{i,k})/\tau\right)} \tag{7}$$

The query-scene combined Gallery Filter Network loss sums positive pair losses over all query-scene pairs:

$$\mathcal{L}^C_{\text{gfn}} = \sum_{i=1}^{N} \sum_{j=1}^{M} \mathcal{I}^Q_{i,j} \ell^C_{i,j} \tag{8}$$

3.2.3 Scene-Only Objective

As a control for the query-scene objective, we also define a simpler objective which uses scene embeddings only, depicted in Figure 3b. This objective attempts to learn the less discriminative concept of whether two scenes share any persons in common, and has the same optimization issue of conflicting attractions and repulsions as the baseline objective. At inference time, it is used in the same way as the other GFN methods.

We define the scene-scene indicator function to denote positive scene-scene pairs as

$$\mathcal{I}^S_{i,j} = \begin{cases} 1 & \text{if } s_i \text{ shares any } q \text{ in common with } s_j \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Similar to Section 3.2.1, we define an index set: $K^S_{i,j} = \{k \in 1, \ldots, M \mid k = j \text{ or } \mathcal{I}^S_{i,j} = 0\}$. Then the loss for a positive scene-scene pair is the cross-entropy loss

$$\ell^S_{i,j} = -\log \frac{\exp\left(\text{sim}(y_i, y_j)/\tau\right)}{\sum_{k \in K^S_{i,j}} \exp\left(\text{sim}(y_i, y_k)/\tau\right)} \tag{10}$$

The scene-only Gallery Filter Network loss sums positive pair losses over all scene-scene pairs:

$$\mathcal{L}^S_{\text{gfn}} = \sum_{i=1}^{M} \sum_{j=1}^{M} [i \neq j] \mathcal{I}^S_{i,j} \ell^S_{i,j} \tag{11}$$

where $[i \neq j]$ is 1 if $i \neq j$ else 0.

3.2.4 Architecture and Optimization

We consider a number of design choices for the architecture and optimization strategy of the GFN to improve its performance.

Architecture. Scene embeddings are extracted in the same way as person embeddings, except that a larger 56×56 pooling size with adaptive max pooling is used vs. the person pooling size of 14×14 with RoI Align. This larger scene pooling size is needed to adequately summarize scene information, since the scene extent is much larger than a typical person bounding box. In addition, the scene `conv5` head and Emb Head are duplicated from the corresponding person modules (no weight-sharing), shown in Figure 2.

Lookup Table. Similar to the methodology used for the OIM objective [35], we use a lookup table (LUT) to store scene and person embeddings from previous batches, refreshing the LUT fully during each epoch. We compare the person and scene embeddings in each batch, which have gradients, with some subset of the embeddings in the LUT, which do not have gradients. Therefore only comparisons of embeddings within the batch, or between the batch and the LUT, have gradients.

Query Prototype Embeddings. Rather than using person embeddings directly from a given batch, we can use the identity prototype embeddings stored in the OIM LUT, similar to [18]. To do so, we lookup the corresponding identity for a given batch person identity in the OIM LUT during training, and substitute that into the objective. In doing so, we discard gradients from batch person embeddings, meaning that we only pass gradients through scene embeddings, and therefore only update the scene embedding module. This choice is examined in an ablation in Section 4.4.

4. Experiments and Analysis

4.1. Datasets and Evaluation

Datasets. For our experiments, we use the two standard person search datasets, CUHK-SYSU [35], and *Person Re-identification in the Wild* (PRW) [40]. CUHK-SYSU comprises a mixture of imagery from hand-held cameras, and shots from movies and TV shows, resulting in significant visual diversity. It contains 18,184 scene images annotated with 96,143 person bounding boxes from tracked (known) and untracked (unknown) persons, with 8,432 known identities. PRW comprises video frames from six surveillance cameras at Tsinghua University in Hong Kong. It contains 11,816 scene images annotated with 43,110 person bounding boxes from known and unknown persons, with 932 known identities.

The standard test retrieval partition for the CUHK-SYSU dataset has 2,900 query persons, with a gallery size of 100 scenes per query. The standard test retrieval partition for the PRW dataset has 2,057 query persons, and uses all 6,112 test scenes in the gallery, excluding the identity. For a more robust analysis, we additionally divide the given train set into separate train and validation sets, further discussed in Supplementary Material Section A.

Evaluation Metrics. As in other works, we use the standard re-id metrics of mean average precision (mAP), and top-1 accuracy (top-1). For detection metrics, we use recall and average precision at 0.5 IoU (Recall, AP).

In addition, we show GFN metrics mAP and top-1, which are computed as metrics of scene retrieval using GFN scores. To calculate these values, we compute the GFN score for each scene, and consider a gallery scene a match to the query if the query person is present in it.

4.2. Implementation Details

We use SGD optimizer with momentum for ResNet models, with starting learning rate 3e-3, and Adam for ConvNeXt models, with starting learning rate 1e-4. We train all models for 30 epochs, reducing the learning rate by a factor of 10 at epochs 15 and 25. Gradients are clipped to norm 10 for all models.

Models are trained on a single Quadro RTX 6000 GPU (24 GB VRAM), and 30 epoch training time using the final model configuration takes 11 hours for the PRW dataset, and 21 hours for the CUHK-SYSU dataset.

Our baseline model used for ablation studies has a ConvNeXt Base backbone, embedding dimension 2,048, scene embedding pool size 56×56, and is trained with 512×512 image crops using the combined cropping strategy (RSC+RFC). It uses the combined prototype feature version of the GFN objective. The final model configuration, used for comparison to other state-of-the-art models, is trained with 640×640 image crops using the altered combined cropping strategy (RSC+RFC2). It uses the combined batch feature version of the GFN objective.

Additional implementation details are given in Supplementary Material Section B.

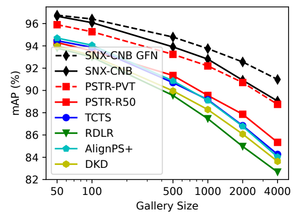

Figure 4: Effect of gallery size on mAP for the CUHK-SYSU dataset. SNX-CNB = SeqNeXt ConvNeXt Base. GFN helps more as gallery size increases.

Occluded	mAP	top-1
SeqNeXt	91.1	89.8
SeqNeXt+GFN	**92.0**	**90.9**

Low-Resolution	mAP	top-1
SeqNeXt	91.4	92.4
SeqNeXt+GFN	**92.0**	**93.1**

Table 1: Performance metrics on two CUHK-SYSU retrieval partitions using either Occluded (top) or Low-Resolution (bottom) query persons.

4.3. Comparison to State-of-the-art

We show a comparison of state-of-the-art methods on the standard benchmarks in Table 2. The GFN benefits all metrics, especially top-1 accuracy for the PRW dataset, which improves by 4.6% for the ResNet50 backbone, and 2.9% for the ConvNeXt Base backbone. Our best model, SeqNeXt+GFN with ConvNext Base, improves mAP by 1.8% on PRW and 1.2% on CUHK-SYSU over the previous best PSTR model. This benefit extends to larger gallery sizes for CUHK-SYSU, shown in Figure 4. In fact, the GFN score-weighting helps more as gallery size increases. This is expected, since the benefit of down-weighting contextually-unlikely scenes, vs. discriminating between persons within a single scene, has a greater effect when there are more scenes compared against.

The GFN benefits CUHK-SYSU retrieval scenarios with occluded or low-resolution query persons, as shown in Table 1. This shows that high quality query person views are not essential to the function of the GFN.

The GFN also benefits both cross-camera and same-camera retrieval, as shown in Table 3. Strong cross-camera performance shows that the GFN can generalize to varying locations, and does not simply pick the scene which is the most visually similar. Strong same-camera performance shows that the GFN is able to use query information, even when all gallery scenes are contextually similar.

To showcase these benefits, we provide some qualitative results in Supplementary Material Section C. These examples show that the GFN uses local person information combined with global context to improve retrieval ranking, even in the presence of difficult confusers.

4.4. Ablation Studies

We conduct a series of ablations using the PRW dataset to show how detection, re-id, and GFN performance are

Method	Backbone	CUHK-SYSU		PRW	
		mAP	top-1	mAP	top-1
Two-step					
IDE [40]	ResNet50	-	-	20.5	48.3
MGTS [5]	VGG16	83.0	83.7	32.6	72.1
CLSA [19]	ResNet50	87.2	88.5	38.7	65.0
IGPN [9]	ResNet50	90.3	91.4	47.2	87.0
RDLR [13]	ResNet50	93.0	94.2	42.9	70.2
TCTS [32]	ResNet50	93.9	95.1	46.8	87.5
End-to-end					
OIM [35]	ResNet50	75.5	78.7	21.3	49.4
IAN [34]	ResNet50	76.3	80.1	23.0	61.9
NPSM [23]	ResNet50	77.9	81.2	24.2	53.1
RCAA [3]	ResNet50	79.3	81.3	-	-
CTXG [37]	ResNet50	84.1	86.5	33.4	73.6
QEEPS [26]	ResNet50	88.9	89.1	37.1	76.7
APNet [41]	ResNet50	88.9	89.3	41.9	81.4
HOIM [4]	ResNet50	89.7	90.8	39.8	80.4
BINet [8]	ResNet50	90.0	90.7	45.3	81.7
NAE+ [6]	ResNet50	92.1	92.9	44.0	81.1
PGSFL [18]	ResNet50	92.3	94.7	44.2	85.2
DKD [39]	ResNet50	93.1	94.2	50.5	87.1
DMRN [14]	ResNet50	93.2	94.2	46.9	83.3
AGWF [12]	ResNet50	93.3	94.2	53.3	87.7
AlignPS [36]	ResNet50	94.0	94.5	46.1	82.1
SeqNet [21]	ResNet50	93.8	94.6	46.7	83.4
SeqNet+CBGM [21]	ResNet50	94.8	95.7	47.6	87.6
COAT [38]	ResNet50	94.2	94.7	53.3	87.4
COAT+CBGM [38]	ResNet50	94.8	95.2	54.0	89.1
MHGAM [20]	ResNet50	94.9	95.9	47.9	88.0
PSTR [2]	ResNet50	94.2	95.2	50.1	87.9
PSTR [2]	PVTv2-B2	95.2	96.2	56.5	89.7
SeqNeXt (ours)	ResNet50	94.1	94.7	50.8	86.0
SeqNeXt+GFN (ours)	ResNet50	94.7	95.3	51.3	90.6
SeqNeXt (ours)	ConvNeXt	96.1	96.5	57.6	89.5
SeqNeXt+GFN (ours)	ConvNeXt	**96.4**	**97.0**	**58.3**	**92.4**

Table 2: Standard performance metrics mAP and top-1 accuracy on the benchmark CUHK-SYSU and PRW datasets are compared for state-of-the-art *two-step* and *end-to-end* models. ConvNeXt backbone = ConvNeXt Base.

Method	Same Cam ID		Cross Cam ID	
	mAP	top-1	mAP	top-1
HOIM [4]	-	-	36.5	65.0
NAE+ [6]	-	-	40.0	67.5
SeqNet [21]	-	-	43.6	68.5
SeqNet+CBGM [21]	-	-	44.3	70.6
AGWF [12]	-	-	48.0	73.2
COAT [38]	-	-	50.9	75.1
COAT+CBGM [38]	-	-	51.7	76.1
SeqNeXt (ours)	82.9	98.5	55.3	80.5
SeqNeXt+GFN (ours)	**85.1**	**98.6**	**56.4**	**82.1**

Table 3: Performance on the PRW test set for query and gallery scenes from the same camera (Same Cam ID) or different cameras (Cross Cam ID).

each impacted by variations in model architecture, data augmentation, and GFN design choices.

In the corresponding metrics tables, we show re-id results by presenting the GFN-modified scores as mAP and top-1, and the difference between unmodified mAP and top-1 with ΔmAP and Δtop-1. This highlights the change in re-id performance specifically from the GFN score-weighting. To indicate the baseline configuration in a table, we use the † symbol, and the final model configuration is highlighted in gray.

	Detection		Re-id				GFN	
GFN Objective	Recall	AP	mAP	top-1	Δ mAP	Δ top-1	mAP	top-1
None	96.0	**93.6**	58.6	88.7	-	-	-	-
Scene-Only	96.0	93.4	56.5	91.9	-0.9	+2.8	16.1	73.3
Base Batch	95.7	93.1	53.9	86.6	-2.6	-2.0	**23.8**	58.4
Base Proto	96.0	**93.6**	55.0	86.2	-3.0	-2.7	22.9	57.8
Comb. Batch	**96.2**	**93.6**	**59.5**	92.2	**+1.1**	+2.9	20.5	**78.8**
Comb. Proto†	96.0	93.4	58.8	**92.3**	**+1.1**	**+3.5**	20.4	78.5

Table 4: Comparison of different options for the GFN optimization objective. "None" does not use the GFN, Scene-Only uses the objective in Section 3.2.3, Base uses the baseline objective in Section 3.2.1, Combined (Comb.) uses the query-scene objective in Section 3.2.2, Batch indicates that batch query embeddings are used, Proto indicates that prototype query embeddings are used. Baseline model is marked with †, final model is highlighted gray.

Results for most of the ablations are shown in Supplementary Material Section D, including model modifications, image augmentation, scene pooling size, embedding dimension, and GFN sampling.

GFN Objective. We analyze the impact of the various GFN objective choices discussed in Section 3.2. Comparisons are shown in Table 4. Most importantly, the re-id mAP performance without the GFN is relatively high, but the re-id top-1 performance is much lower than the best GFN methods. Conversely, the Scene-Only method achieves competitive re-id top-1 performance, but reduced re-id mAP.

The Base methods were found to be significantly worse than all other methods, with GFN score-weighting actually reducing GFN performance. The Combined methods were the most effective, better than the Base and Scene-Only methods for both re-id and GFN-only stats, showcasing the improvements discussed in Section 3.2.2. In addition, the success of the Combined objective can be explained by two factors: 1) similarity relationship between scene embeddings and 2) query information given by query-scene embeddings. The Scene-Only objective, which uses only similarity between scene embeddings, is functional but not as effective as the Combined objective, which uses both scene similarity and query information. Since the Scene-Only objective incorporates background information, and does not use query information, we reason that the provided additional benefit of the Combined objective comes from the described mechanism of query excitation of scene features, and not from *e.g.*, simple matching of the query background with the gallery scene image.

Finally, the Batch and Proto modifiers to the Combined and Base methods were found to be relatively similar in performance. Since the Proto method is simpler and more efficient, we use it for the baseline model configuration.

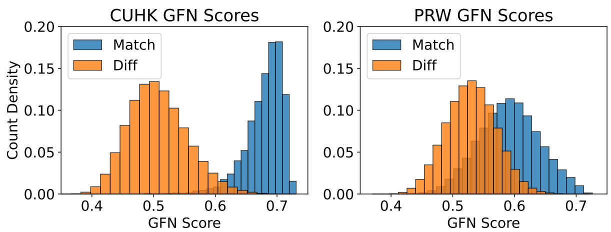

Figure 5: GFN score histograms for the CUHK-SYSU and PRW test sets. Matches and non-matches (Diffs) are shown for queries in the gallery size 4,000 set for CUHK-SYSU, and the full gallery for PRW.

4.5. Filtering Analysis

GFN Score Threshold. We consider selection of the GFN score threshold value to use for filtering out gallery scenes during retrieval. In Figure 5, we show histograms of GFN scores for both CUHK-SYSU and PRW. We introduce another metric to help analyze computation savings from the filtering operation: the fraction of negative gallery scenes which can be filtered out (negative predictive value) when using a threshold which keeps 99% of positive gallery scenes (recall). For the histograms shown, this value is 91.4% for CUHK-SYSU, and only 11.5% for PRW.

In short, this is because there is greater variation in scene appearance in CUHK-SYSU than PRW. This results in most query-gallery comparisons for CUHK-SYSU evaluation occurring between scenes from clearly different environments (e.g., two different movies). While the GFN score-weighting improves performance for both same-camera and cross-camera retrieval, shown in Table 3, query-scene scores used for hard thresholding may be less discriminative for nearly-identical scenes as in PRW vs. CUHK-SYSU, shown in Figure 5. Still, the GFN top-1 score for the final PRW model was 78.4%, meaning that 78.4% of queries resulted in the correct gallery scene being ranked first using only the GFN score.

Compute Cost. In Table 5, we show the breakdown of percent time spent on shared computation, GFN-only computation, and detector-only computation. Since most computation time ($\sim$60%) is spent on detection, with only ($\sim$5%) of time spent on GFN-related tasks, there is a large cost savings from using the GFN to avoid detection by filtering gallery scenes. Exactly how much time is saved in practice depends on the relative number of queries vs. the gallery size, and how densely populated the gallery scenes are with persons of interest.

To give an understanding of compute savings for a single query, we show some example calculations using the conservative recall requirement of 99%. For CUHK-SYSU, we have 99.9% of gallery scenes negative, 91.4% of negative gallery scenes filtered, and 61.0% of time spent doing detection on gallery scenes, resulting in 55.7% computation saved using the GFN compared to the same model without the GFN. For PRW, the same calculation yields 6.6% computation saved using the GFN.

	Shared		**GFN**		**Detection**	
	Backbone	Query Emb.	Scene Emb.	GFN Scores	RPN	R-CNN($\times$2)
CUHK Time (%)	33.7	<0.1	5.3	<0.1	19.2	41.8
	33.7		5.3		61.0	
PRW Time (%)	36.9	<0.1	5.3	<0.1	16.1	41.7
	36.9		5.3		57.8	

Table 5: Percent computation time averaged per query of shared feature extraction, GFN, and detection on the CUHK-SYSU (gallery size 4,000) and PRW (gallery size full) test sets.

5. Conclusion

We describe and demonstrate the Gallery Filter Network, a novel module for improving accuracy and efficiency of person search models. We show that the GFN can efficiently filter gallery scenes under certain conditions, and that it benefits scoring for detects in scenes which are not filtered. We show that the GFN is robust under a range of different conditions by testing on different retrieval sets, including cross-camera, occluded, and low-resolution scenarios. In addition, we show that the benefit given by GFN score-weighting increases as gallery size increases.

Separately, we develop the base SeqNeXt person search model, which has significant performance gains over the original SeqNet model. We offer a corresponding training recipe to train efficiently with improved regularization, using an aggressive cropping strategy. Taken together, the SeqNeXt+GFN combination yields a significant improvement over other state-of-the-art methods. Finally, we note that the GFN is not specific to SeqNeXt, and can be easily combined with other person search models.

Societal Impact. It is important to consider the potential negative impact of person search models, since they are ready-made for surveillance applications. This is highlighted by the PRW dataset being entirely composed of surveillance imagery, and the CUHK-SYSU dataset containing many street-view images of pedestrians.

We consider two potential advantages of advancing person search research, and doing so in an open format. First, that person search models can be used for beneficial applications, including aiding in finding missing persons, and for newly-emerging autonomous systems that interact with humans, *e.g.*, automated vehicles. Second, it allows the research community to understand how the models work at a granular level, and therefore benefits the potential for counteracting negative uses when the technology is abused.

Acknowledgements

The authors would like to thank Wesam Sakla and Michael Goldman for helpful discussions and feedback.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving Into High Quality Object Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, June 2018. ISSN: 2575-7075.

[2] Jiale Cao, Yanwei Pang, Rao Muhammad Anwer, Hisham Cholakkal, Jin Xie, Mubarak Shah, and Fahad Shahbaz Khan. PSTR: End-to-End One-Step Person Search With Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9458–9467, 2022.

[3] 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 Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018*, volume 11213, pages 86–102. Springer International Publishing, Cham, 2018. Series Title: Lecture Notes in Computer Science.

[4] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Bernt Schiele. Hierarchical Online Instance Matching for Person Search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):10518–10525, Apr. 2020. Number: 07.

[5] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person Search via a Mask-Guided Two-Stream CNN Model. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018*, Lecture Notes in Computer Science, pages 764–781, Cham, 2018. Springer International Publishing.

[6] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. Norm-Aware Embedding for Efficient Person Search. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12612–12621, Seattle, WA, USA, June 2020. IEEE.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR, Nov. 2020. ISSN: 2640-3498.

[8] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Bi-Directional Interaction Network for Person Search. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2836–2845, Seattle, WA, USA, June 2020. IEEE.

[9] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Instance Guided Proposal Network for Person Search. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2582–2591, June 2020. ISSN: 2575-7075.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. arXiv:2010.11929 [cs].

[11] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality Reduction by Learning an Invariant Mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, June 2006. ISSN: 1063-6919.

[12] Byeong-Ju Han, Kuhyeun Ko, and Jae-Young Sim. End-to-End Trainable Trident Person Search Network Using Adaptive Gradient Propagation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 905–913, Oct. 2021. ISSN: 2380-7504.

[13] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. Re-ID Driven Localization Refinement for Person Search. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9813–9822, Oct. 2019. ISSN: 2380-7504.

[14] Chuchu Han, Zhedong Zheng, Changxin Gao, Nong Sang, and Yi Yang. Decoupled and Memory-Reinforced Networks: Towards Effective Feature Learning for One-Step Person Search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1505–1512, May 2021. Number: 2.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):386–397, Feb. 2020. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE.

[17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, June 2018. ISSN: 2575-7075.

[18] Hanjae Kim, Sunghun Joung, Ig-Jae Kim, and Kwanghoon Sohn. Prototype-Guided Saliency Feature Learning for Person Search. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4863–4872, June 2021. ISSN: 2575-7075.

[19] Xu Lan, Xiatian Zhu, and Shaogang Gong. Person Search by Multi-Scale Matching. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018*, Lecture Notes in Computer Science, pages 553–569, Cham, 2018. Springer International Publishing.

[20] Yang Li, Huahu Xu, Minjie Bian, and Junsheng Xiao. Cross-scale global attention feature pyramid network for person search. *Image and Vision Computing*, 116:104332, Dec. 2021.

[21] Zhengjia Li and Duoqian Miao. Sequential End-to-end Network for Efficient Person Search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):2011–2019, May 2021. Number: 3.

[22] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, July 2017. ISSN: 1063-6919.

[23] Hao Liu, Jiashi Feng, Zequn Jie, Karlekar Jayashree, Bo Zhao, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. Neural Person Search Machines. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 493–501, Oct. 2017. ISSN: 2380-7504.

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Oct. 2021. ISSN: 2380-7504.

[25] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[26] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-Guided End-To-End Person Search. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 811–820, Long Beach, CA, USA, June 2019. IEEE.

[27] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, Jan. 2019. arXiv:1807.03748 [cs, stat].

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015. ISSN: 1063-6919.

[30] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully Convolutional One-Stage Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9626–9635, Oct. 2019. ISSN: 2380-7504.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[32] Cheng Wang, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. TCTS: A Task-Consistent Two-Stage Framework for Person Search. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11949–11958, Seattle, WA, USA, June 2020. IEEE.

[33] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVT v2: Improved baselines with Pyramid Vision Transformer. *Computational Visual Media*, 8(3):415–424, Sept. 2022.

[34] 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, Mar. 2019.

[35] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint Detection and Identification Feature Learning for Person Search. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3376–3385, Honolulu, HI, July 2017. IEEE.

[36] Yichao Yan, Jinpeng Li, Jie Qin, Song Bai, Shengcai Liao, Li Liu, Fan Zhu, and Ling Shao. Anchor-Free Person Search. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7686–7695, June 2021. ISSN: 2575-7075.

[37] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. Learning Context Graph for Person Search. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2153–2162, Long Beach, CA, USA, June 2019. IEEE.

[38] Rui Yu, Dawei Du, Rodney LaLonde, Daniel Davila, Christopher Funk, Anthony Hoogs, and Brian Clipp. Cascade Transformers for End-to-End Person Search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7267–7276, 2022.

[39] Xinyu Zhang, Xinlong Wang, Jia-Wang Bian, Chunhua Shen, and Mingyu You. Diverse Knowledge Distillation for End-to-End Person Search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3412–3420, May 2021. Number: 4.

[40] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person Re-identification in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3346–3355, Honolulu, HI, July 2017. IEEE.

[41] Yingji Zhong, Xiaoyu Wang, and Shiliang Zhang. Robust Partial Matching for Person Search in the Wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6826–6834, Seattle, WA, USA, June 2020. IEEE.