Journal Pre-proof

Query-Guided Networks for Few-shot Fine-grained Classification and Person Search

Bharti Munjal, Alessandro Flaborea, Sikandar Amin,
Federico Tombari, Fabio Galasso

PII: S0031-3203(22)00529-5
DOI: https://doi.org/10.1016/j.patcog.2022.109049
Reference: PR 109049

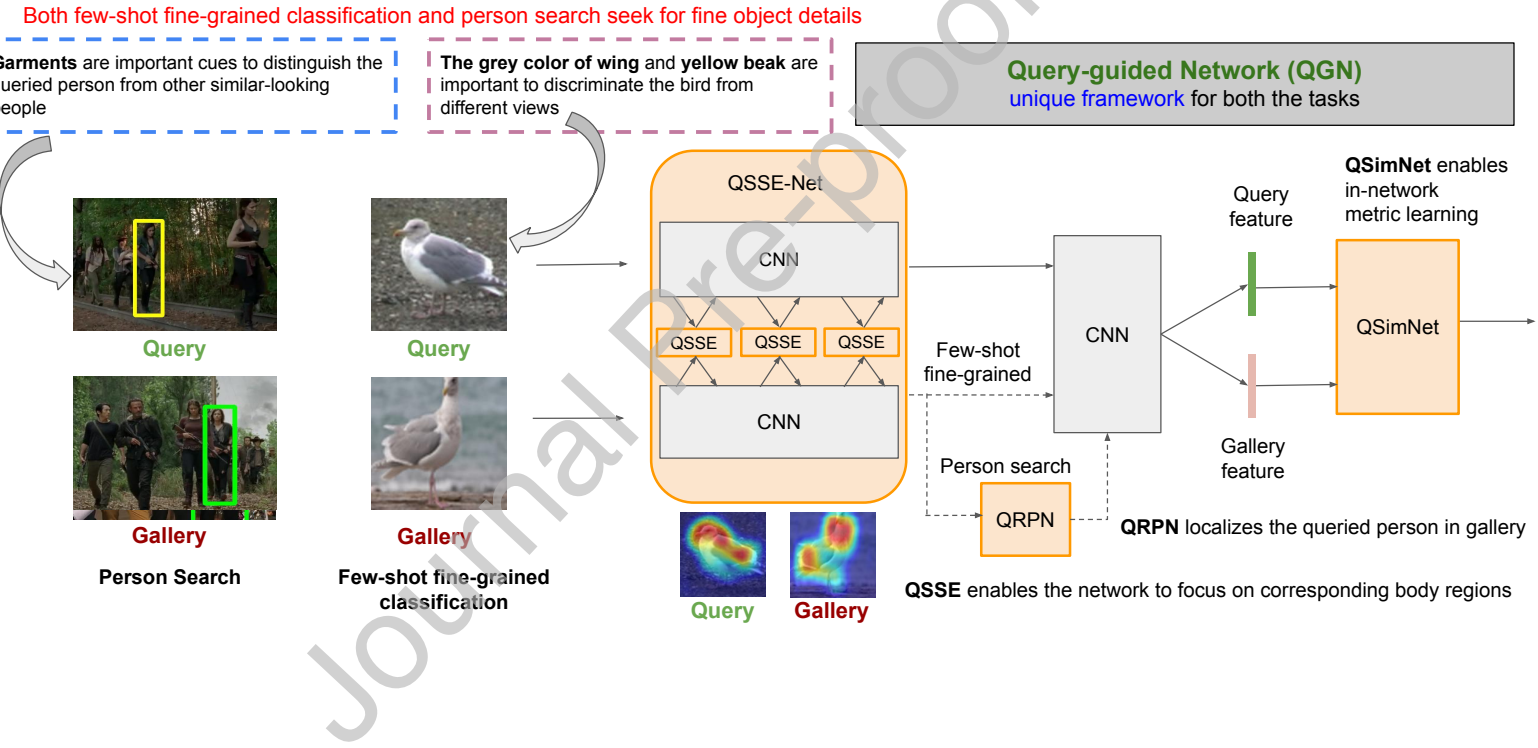
To appear in: *Pattern Recognition*

Received date: 29 September 2021
Revised date: 2 August 2022
Accepted date: 20 September 2022

Please cite this article as: Bharti Munjal, Alessandro Flaborea, Sikandar Amin, Federico Tombari, Fabio Galasso, Query-Guided Networks for Few-shot Fine-grained Classification and Person Search, *Pattern Recognition* (2022), doi: https://doi.org/10.1016/j.patcog.2022.109049

Highlights

1. Prior works treat few-shot fine-grained classification and person search separately
2. The proposed Query-guided Networks (QGN) address both tasks in a unified framework
3. QGN introduces query-guidance for the task of few-shot fine-grained recognition
4. State-of-the-art performance on CUB, FGVC-Aircraft, Stanford Dogs datasets
5. The paper performs an in-depth analysis of each of the query-guided components of QGN



Query-Guided Networks for Few-shot Fine-grained Classification and Person Search

Bharti Munjal[a,*], Alessandro Flaborea[b], Sikandar Amin[c], Federico Tombari[a,d], Fabio Galasso[b]

[a]*Department of Informatics, Technical University of Munich, Germany*
[b]*Department of Computer Science, Sapienza University of Rome, Italy*
[c]*Magic Leap Zurich, Switzerland*
[d]*Google Zurich, Switzerland*

Abstract

Few-shot fine-grained classification and person search appear as distinct tasks and literature has treated them separately. But a closer look unveils important similarities: both tasks target categories that can only be discriminated by specific object details; and the relevant models should generalize to new categories, not seen during training.

We propose a novel unified Query-Guided Network (QGN) applicable to both tasks. QGN consists of a Query-guided Siamese-Squeeze-and-Excitation subnetwork which re-weights both the query and gallery features across all network layers, a Query-guided Region Proposal subnetwork for query-specific localisation, and a Query-guided Similarity subnetwork for metric learning.

QGN improves on a few recent few-shot fine-grained datasets, outperforming other techniques on CUB by a large margin. QGN also performs competitively on the person search CUHK-SYSU and PRW datasets, where we perform in-depth analysis.

Keywords: Meta-Learning, Few-shot Learning, Fine-grained Classification, Person Search, Person Re-Identification

*Corresponding author
Email address: munjalbharti@gmail.com (Bharti Munjal)

1. Introduction

Few-shot fine-grained classification and person search share important similarities, as they both require paying attention to the details, e.g. what distinguishes a person from other people, or a bird from other possibly similar races. Both fields have progressed largely in recent years [1, 2]. Few-shot learning eases the burden of large data collections when generalizing to new unseen (possibly rare) classes. Person search is useful for video surveillance, long term tracking and person verification. Both tasks face the similar challenges of background clutter, illumination and viewpoint changes, occlusions, image blur and distortions, including non-rigid deformations of the object body pose [3, 4].

Person search is the task of finding a specific person, as provided by a single query image, within a gallery image. It consists of localization within the gallery (detection) and re-identification (classification based on the single query example). Few-shot learning similarly stands for recognizing the queried object, either classifying or detecting, typically from a single or multiple (i.e., five) examples (1- and 5-shot learning). Fine-grained classification specifically describes the challenge of recognizing an object (bird, aircraft, dog etc.) from a few details (the shape of the beak, the pattern on the wings etc.). Person search is therefore a one-shot fine-grained classification task, which includes detection. Note that in few-shot fine-grained classification the *query-gallery* pair is termed *support-query* respectively, especially confusing for the role of the query. Throughout this work, we adopt the person search terminology and search a *query* person or object within a *gallery* image. See Sec. 3 for more details.

We propose a novel unified Query-Guided Network (QGN) to address both person search and few-shot fine-grained classification. Query guidance is novel and stands for processing the query and gallery images jointly, with a Siamese network design and query-gallery interaction modules. By contrast, prior literature in person search [3, 5, 6] and few-shot learning [7, 8, 9] typically extracts separate features for the query and gallery images, which prevents their models from emphasizing query-specific patterns in the gallery search.

QGN proposes three query-gallery interaction modules: **i.** the Query-guided Siamese Squeeze-and-Excitation Network (QSSE) re-weights both the query and gallery channel features, jointly conditioned on both images; **ii.** the Query Similarity Network (QSimNet) learns a similarity metric which is specific for comparing with the query; **iii.** the Query-guided RPN (QRPN) is used for detection, to provide query-specific proposals (besides the classic RPN).

The modularity of QGN allows to evaluate the core idea of extensively using query guidance in retrieval for detection and classification tasks. In both cases, query guidance enhances the relevance of ID features in the network backbone, matching function and, if present, in the region proposal. We consider person search as the detection task (in any case, this subsumes person re-identification) and few-shot fine-grained recognition as the classification task (to the best of our knowledge, there is no established few-shot fine-grained object detection benchmark yet).

Query-guidance is novel in the few-shot context. We evaluate QGN on five-widely adopted few-shot fine-grained datasets: CUB [10], Stanford Cars [11], FGVC-Aircraft [12], Stanford Dogs [13], and Oxford Flowers [14]. QGN achieves state-of-the-art results on CUB, FGVC-Aircraft and Stanford Dogs. Particularly on CUB, QGN surpasses the current best S2M2 [8] by a large margin, i.e. 12pp and 5pp in 1- and 5-shot learning experiments, respectively. Moreover, when employing a shallower ResNet18, the performance of QGN surpasses S2M2, which employs the deeper WRN [8], by 3.1pp for 1-shot learning.

For person search, we add our query-guided components on top of a recently improved OIM implementation [1], and achieve competitive performance with the state of the art on the large scale CUHK-SYSU [3] and PRW [15] datasets. We report comparison with several competing person search techniques, including the ones following our original work [16]. Both in person search and in few-shot fine-grained classification, we perform an in-depth analysis, including diverse backbones (ResNet10, ResNet18, ResNet50, WRN-28-10). Furthermore, we demonstrate the intuition of our proposed query-guided components via qualitative visualizations on both tasks.

2. Related Work

We review prior art on few-shot learning, fine-grained classification and person search, emphasizing methods which condition the feature extraction upon the query. To the best of our knowledge, QGN is the first technique addressing both tasks and it is the first query-guidance approach for few-shot fine-grained classification.

Few-shot learning. Few-shot learning aims to train models that can rapidly adapt and generalize to new concepts using only a few samples. The copious recent progress in the field can be loosely divided into five categories. In the first, *metric-based* methods [17, 18] learn a shared embedding space for the comparison of the feature embeddings from the query and the gallery images. The proposed QSimNet resembles the relation module in the Relation Network [17], but the input features of query and gallery are jointly extracted and end-to-end trained. In the second category, *optimization based* methods [19] adjust the optimization algorithm to learn from a few examples. Here the most popular is MAML [19], which optimizes the initialization of the gradient-descent-based learner. *Data hallucination* may be a third direction, based on the data augmentation and the scarce provided data.

More recently, [9] proposed a simpler *transfer learning* approach using a distance-based classifier, which is competitive with other more sophisticated approaches. S2M2 [8] extends their work with self-supervision techniques [20]. Following [9, 8], QGN also employs the *non-episodic* training, hence it does not need to train separately for different few-shot protocols. Unlike transfer learning methods, QGN jointly processes the query and the gallery with a Siamese network model and it does not need any fine-tuning at inference time.

Finally, the category of *dynamic network conditioning* methods uses the query or gallery examples to either tune or condition the network by *attention* based mechanism [21] or *generate network parameters* [22]. Matching networks [23] apply conditioning as post-processing with a bidirectional LSTM. [24] uses a weight-centric learning strategy to push samples closer to their corre-

sponding classifier weights. Other approaches generate weights by means of kernel generator or by combining basis convolutional kernel filters [22]. These techniques relate to QSSE, which we employ for feature extraction, however our approach is the sole to make use of both the query and gallery features from the very first layers. Similar to ours, CAM [21] generates query-gallery cross-attention maps, but it focuses on image parts, rather than entire feature channels, as we do. Also, the correlation layer of CAM is applied only once at the output layer, due to its high memory and runtime requirements, while our simpler QGN is applied at all network layers, which results in the query-gallery interaction across both coarser and finer details.

Few-shot fine-grained classification. Fine-grained differs mainly from general few-shot learning as it focuses on categories with subtle distinctive traits, e.g. species of birds, dogs, flowers, car models. This is more complex and less researched. Within this literature, [4] targets fine-grained few-shot recognition by learning pose normalized embedding and uses extra part annotations. [25] uses attention modules after the feature extractor to infer spatial and channel attentions. [26] employs a multi-scale feature pyramid and a multi-level attention pyramid to extract features of different granularities. More recently, [9] evaluates the generic few-shot methods including ProtoNet [7], MatchingNet [23], RelationNet [17] and MAML [19] on few-shot fine-grained classification. S2M2 [8] also evaluates its approach on the fine-grained case. [27] propose a unifying loss for various fine-grained tasks. Unlike the above methods, our QGN is a Siamese model and it leverages query-gallery cross-attention.

Person Search. There are several person search techniques but they are distinct from the previous, as no methods address both tasks. In person search, we distinguish sequential methods [6, 28], which cascade the person detection and person re-identification sub-tasks, from joint methods [29, 30], which perform both sub-tasks with a single network. The latter lag a bit behind sequential models in performance and are more complex to train, since detection and re-identification are conflicting sub-tasks. However these require in general less memory and computational resources, and are therefore preferable for indus-

trial applications. QGN belongs to this second category, but the proposed query-guidance components are applicable to a sequential method, too.

Among the joint models, QGN relates to [3] which introduces Online Instance Matching (OIM) into Faster RCNN [31] as an additional multi-task loss. OIM is the de-facto standard re-identification loss, adopted by most recent person search approaches [1, 32] as well as by QGN. PGA [32] uses the class prototype as a guidance for person attention. AlignPS [1] proposes an anchor free framework for person search with a feature aggregation module. Similarly, BINet [33] and NAE [5] build on top of OIM. BINet [33] employs an additional parallel branch that takes cropped patches and supervises the joint model with interaction losses. NAE [5] decomposes the embeddings of OIM into angle and norm to accomplish re-ID and detection respectively. [34] uses a hierarchical distillation strategy to transfer knowledge from a stronger teacher model to a student model. QGN is the first to introduce query-gallery interaction modules at different stages of the network, as well as throughout the backbone.

Query-guided person search. Prior work from ours [16] was the first to introduce query guidance for person search. Afterwards, this has been adopted by a few techniques, including TCTS [35] and IGPN [6]. TCTS proposes an identity-guided query detector to produce query-like person boxes for the subsequent re-ID network. IGPN replaces the standard two-stage detector with a query- or instance-guided detector. IGPN adopts the Siamese RPN which correlates the query and gallery feature maps. By contrast, the proposed QRPN takes the query image crop at the input and re-weights the feature channels of the gallery image, emphasizing the traits of the person which we are searching for. Also, both IGPN and TCTS are sequential approaches that use two different models for detection and re-identification, while ours is a joint approach. Note that the joint models require less resources as compared to the sequential approaches as both the model parameters and processing are shared by the backbone. Additionally, learning joint models provides an appealing multi-task objective and addressing this successfully may result in a better use of data, higher performance and a better direction towards general intelligence, i.e. net-

works which understand multiple aspects of the scene.

3. Method

In this section, we first formulate few-shot fine-grained classification and person search tasks. Then we discuss the proposed model and the three query-guided modules, as well as the optimization details.

3.1. Problem formulation

Let us describe the few-shot fine-grained classification and the person search tasks in a unified way.

The training and test sets for both tasks can be given as $D_{train} = \{(x_i, y_i)\}_{i=1}^{N_{train}}$ with C_{train} classes and $D_{test} = \{(x_i, y_i)\}_{i=1}^{N_{test}}$ with C_{test} classes, respectively. Here, x_i represents the images and y_i their corresponding ground-truth annotations. In particular, y_i stands for the object classes in the case of few-shot fine-grained classification; and it means the person-ID and its location in the image for the task of person search. The set of C_{train} and C_{test} classes in D_{train} and D_{test} are disjoint, i.e. at test time the model needs to classify new classes and person-IDs.

Following literature from both tasks, we employ an episodic evaluation protocol, where a subset D_{novel} is sampled from D_{test} with C_{novel} novel classes in each episode. A part of D_{novel}, i.e. K examples from each of the C_{novel} classes, is considered as query. The remaining part of D_{novel} is the gallery, where the model needs to find the queries.

In the **few-shot** case, $(K + L)C_{novel}$ examples are sampled per episode as D_{novel} , i.e. C_{novel} classes with K examples per class as query. This is termed C_{novel}-way K-shot classification. While another L examples per class are used as gallery. C_{novel} also represents the complexity of the evaluation. Larger C_{novel} means more competition among classes during classification. On the other hand, K represents the number of examples per class in C_{novel} that we can use as query. Larger K means more information per class. Typically, K is either 1 (1-shot learning) or 5 (5-shot learning).

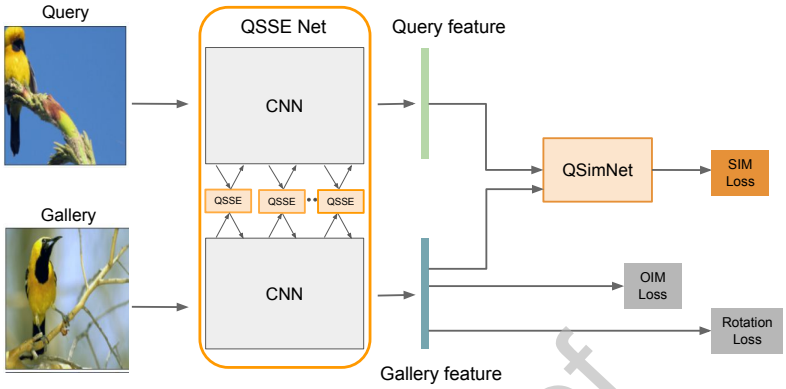


Figure 1: Our proposed query-guided network for few-shot fine-grained classification. The *bottom network* OIM [3] with auxiliary rotation loss is our baseline OIM_R. We pair the baseline network with a siamese branch on top that takes a *query* and guides the *bottom network* at different levels. CNN here represents standard network architecture (ResNet10, ResNet18, WRN) followed by global average pooling. Note that we follow the person search terminology here: *query* refers to the example for which we already know the class and *gallery* needs to be classified. Our proposed query-guidance blocks are given in orange.

In **person search**, an episode D_{novel} is sampled per query example. Here, D_{novel} includes all positive examples corresponding to that query and a large number of random negatives from D_{test}, e.g. for CUHK-SYSU [3] the size of D_{novel} is typically 101(= 100 gallery +1 query). Therefore, $C_{novel} = 2$ and $K = 1$. $C_{novel} = 2$ means person search follows a binary classification strategy i.e. either the gallery sample matches the query or not. $K = 1$ means only one example per class is given as a query at one time. Therefore, person search can be viewed as a special case of few-shot classification, i.e. 1-shot learning.

Note: The terminology used in few-shot classification literature is different from that of person search. In person search, the query image is the one for which the class (or ID) is already known, while the gallery image needs to be classified. Whereas, in few-shot classification, the query is the image that needs to be classified and the support is the image for which the class is already known. To keep the terminology consistent, we adopt the *query-gallery* convention of person search for few-shot case as well.

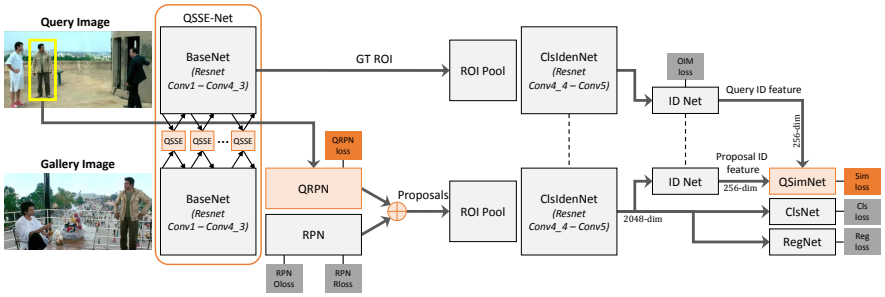


Figure 2: Our proposed query guided network architecture for person search. We pair the reference OIM [3] *bottom network* with a novel Siamese *top network*, to process the query and guide the person search at different levels of supervision (cf. Sec. 3). The novel query-guidance blocks of our approach, displayed in orange, are trained end-to-end with the whole network with specific loss functions (*darker orange boxes*).

3.2. Query-guided Networks

When provided with one or a few query samples, humans focus on its relevant and distinguishing features to find a corresponding gallery image and the object within it. Inspired by this, QGN proposes to process jointly the query and gallery images by a Siamese network design, and to model the query-gallery interactions by query-guided modules.

Few-shot fine-grained classification is accomplished by a Siamese network which processes the query and gallery images together, to produce an embedding for each of them, which is used to classify the gallery class to one of the novel classes in D_{novel}. The relevant overall QGN model is illustrated in Figure 1. The image embeddings are computed by two convolutional backbones. QGN contributes several Query-guided Siamese Squeeze-and-Excitation Network (QSSE) blocks, which relate the feature extraction at multiple layers of the backbone. Finally QGN realizes the classification of the embeddings by a Query Similarity Network (QSimNet), which learns the final metric similarity score. These components are described in detail in Sec.3.3. The implementation of each branch in the Siamese network draws details from [8] and leverages for training the OIM loss [3].

Person search is realized by two parallel Siamese detection networks, which extract the object crops from the query and gallery images, computes an embedding and compares those to assess whether they contain the same or different classes. The proposed QGN model is illustrated in Fig. 2. The image embeddings are extracted with convolutional backbones, leveraging the multi-layers query-gallery interaction by the QSSE. Then the object crops are extracted from the gallery by the proposed Query-guided Region Proposal Network (QRPN), i.e. proposals for bounding boxes tailored to the queried object, which integrates the proposals of a standard RPN [31]. The top proposals are then passed to the subsequent network with a multi-task head for classification (person vs non-person), localization refinement (regression offsets), and ID feature learning. Finally, the ID embeddings of query and each gallery proposal are compared by the QSimNet to distinguish same Vs different IDs. Details for the QGN components are provided in Sec. 3.3.

The implementation of each detection parallel branch follows details of [3], including the OIM loss. Differently from the few-shot fine-grained, person search includes a detection task, so the entire query and gallery images are provided to the network, not just the person crops. Note that we do not need proposals for the query branch, since the query crop is given as input.

3.3. Query-guided Network Components

We propose three components to provide query-guidance at different stages of the Siamese networks. QSSE considers joint global context of the query and gallery to re-calibrate the channel features of the convolutional backbones. QRPN generates query-like proposals exploiting the query-crop specific patterns. QSimNet learns a distance metric to compare the query- gallery features.

In person search (Fig. 2), we adopt all three components. In few-shot fine grained classification (Fig. 1), there is no need to generate candidate proposals and QGN consists only of QSSE and QSimNet. In both cases, all network parts are trained end-to-end.

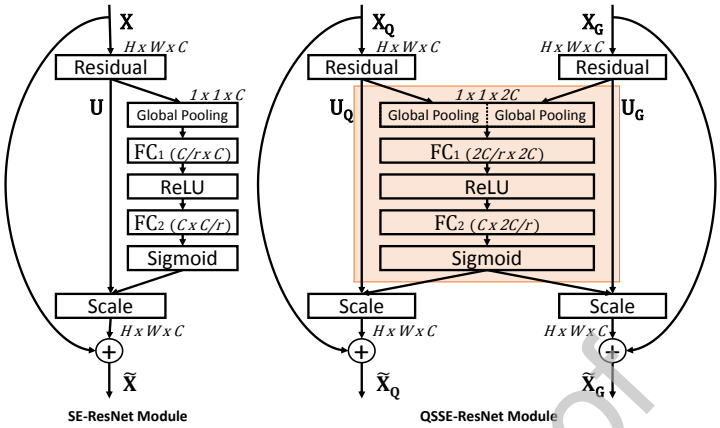


Figure 3: On the left a standard SE block [36] is shown. On the right is our proposed Query-guided Siamese Squeeze-and-Excitation Network (QSSE-Net). The globally-pooled query and gallery features after the ResNet block are concatenated and jointly used to re-calibrate feature channels of both query and gallery. This way QSSE considers both intra- and inter-channel dependencies.

3.3.1. Query-guided Siamese Squeeze-and-Excitation Network (QSSE)

The query and gallery objects in the images may be taken from different viewpoints and with different lighting conditions. Their embeddings should ideally disentangle these nuances. To this goal, we propose the QSSE module, which leverages the interaction of query and gallery. More specifically, as shown in Fig. 2, the QSSE modules, inserted at the output of each network block (e.g. residual block for ResNet), allow a joint re-calibration of the feature maps.

The QSSE module draws inspiration from SE-Net [36], extending it to pairs of images (Fig. 3). In more detail, inside a QSSE, first a *squeeze* operation is performed by global average pooling of query and gallery features. This operation summarizes the spatial information of each of the C channels, giving descriptors $\mathbf{z}_q$ and $\mathbf{z}_g \in \mathbb{R}^C$ for query and gallery respectively.

After this, an *excitation* operation is performed where the two descriptors are first concatenated $[\mathbf{z}_q, \mathbf{z}_g] \in \mathbb{R}^{2C}$ and then passed through a non-linear bottleneck. The first layer FC_1 of the bottleneck is for dimensionality reduction, shrinking the dimension of the concatenated descriptor by a factor of r. This

reduced feature ($\frac{2C}{r}$) is then passed through the ReLU operation (δ) modeling non-linear dependencies between channels. Finally, the feature is expanded to C dimensions by the next fully connected layer FC_2, followed by sigmoid activation (σ) to generate the weight vector s $\in \mathbb{R}^C$. Mathematically, the Siamese squeeze-and-excitation operation is given by

$$\mathbf{s} = \mathrm{F}_{ex}(\mathbf{z}_q, \mathbf{z}_g; \mathbf{W}) = \sigma(\ \mathbf{W}_2\ \delta(\ \mathbf{W}_1[\mathbf{z}_q, \mathbf{z}_g]\ )\ ) \tag{1}$$

where the parameters of the first and second fully connected layers are, respectively, $\mathbf{W}_1 \in \mathbb{R}^{\frac{2C}{r}\mathrm{x}2C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C\mathrm{x}\frac{2C}{r}}$.

Following [36], we set the reduction ratio r to 16 in all our experiments. As shown in Fig. 3, the *scale* operation employs the weight vector s to re-weight the residual outputs $\mathbf{U_Q}$ (for query) and $\mathbf{U_G}$ (for gallery), by channel-wise multiplication. These scaled outputs are then added to the original features $\mathbf{X}_\mathrm{Q}$ and $\mathbf{X}_\mathrm{G}$ via *skip connections*, giving outputs $\widetilde{\mathbf{X}}_\mathrm{Q}$ and $\widetilde{\mathbf{X}}_\mathrm{G}$ respectively. Mathematically, the above operation is defined as

$$\begin{aligned} \widetilde{\mathbf{X}}_\mathrm{Q} &= \mathbf{X_Q} + \mathbf{s} \odot \mathbf{U_Q} \\ \widetilde{\mathbf{X}}_\mathrm{G} &= \mathbf{X_G} + \mathbf{s} \odot \mathbf{U_G} \end{aligned} \tag{2}$$

where $\odot$ denotes the *channel-wise* scaling operation.

3.3.2. Query-guided RPN (QRPN)

QRPN is an attention-based region proposal network that leverages the local query features to generate query-like object proposals. QRPN consists of a channel-wise attention sub-network followed by a standard RPN [31], as shown in Fig. 4. The attention network uses the cropped query features to re-weight the feature channels of the gallery image. The re-weighted features are then passed to a standard RPN to generate object proposals.

In more detail, the query-crop features are first pooled using a ROI-pool [31]. We then pass the pooled query features to a non-linear bottleneck. The first layer FC_1 of the bottleneck reduces the pooled features to $\mathbb{R}^{C/r}$, where $C = 1024$ and $r = 16$. Note that FC_1 is applied to all pixels of all the channels of

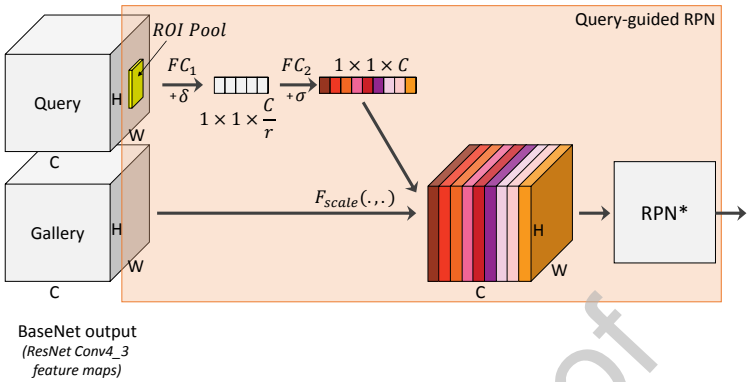


Figure 4: Query-guided Region Proposal Network (QRPN) adapts squeeze-and-excitation to generate weights from local query and re-calibrates gallery feature channels. The re-weighted gallery features are then passed to RPN* where RPN* is the standard RPN but does not compute regression offsets.

the pooled map. In this way, our attention mechanism leverages the spatially *localized* query crop patterns to emphasize particular gallery channels. This also gives the network layer more freedom and lets the optimization dictate what specific local patterns to highlight, instead of just global features. This is in contrast with the *squeeze* operation of SE-Net [36]. The second fully connected layer FC_2 then expands the features back to C dimensions, followed by a sigmoid (σ) activation to generate weights. Finally, the output weights are used to re-calibrate the gallery features and not the query itself.

We further complement QRPN with the standard RPN in a parallel branch, that takes into account generic objectness score (cf. Fig 2). This helps in retrieving further proposals when they are quite different from the query. The objectness score from RPN and query-similarity score from QRPN are summed up to generate final score for each anchor which is used for non-maximal suppression (NMS) at the stage of proposal generation. Note that both RPN included in QRPN and the parallel RPN follow the same design and use same anchors.

QRPN is trained using **QRPN loss** which is a binary cross-entropy loss

given as,

$$L_{qrpn} = -\frac{1}{N}\sum_{N}\log(p_n^u) \tag{3}$$

where p_n^u is the probability of the true class u for the n^{th} anchor out of a total of N anchors.

3.3.3. Query-guided Similarity Net (QSimNet)

QSimNet is a deep query-dependent metric that is trained end-to-end with other network components. Unlike standard offline metrics such as the euclidean distance [3, 5], QSimNet alters the similarity measures for each query, to account for the relative importance of attributes such as e.g. color and shape.

As shown in Fig. 5, QSimNet works by first calculating the L2 distance between the two features, i.e element-wise subtraction and square operation. This is followed by batch normalization and a fully connected layer with two outputs. Finally, a softmax is applied to generate similarity/dissimilarity scores.

QSimNet is trained using **Sim loss** L_{sim} which is defined as the binary cross-entropy loss similar to L_{qrpn}. L_{sim} is given as,

$$L_{sim} = -\frac{1}{N}\sum_{N}\log(p_n^t) \tag{4}$$

where N defines the number of pairs in the mini-batch and p_n^t is the probability of the true class t for the n^{th} pair.

3.4. Training Query-guided Networks

We discuss in details the optimization of QGN for each of the task.

3.4.1. Few-shot fine-grained classification

The QGN network is optimized in an end-to-end fashion, which considers both the classification backbone, as well as the QSSE and QSimNet.

Self-supervision has been proven to improve few-shot learning in various recent works [20, 8] as it helps to overcome *supervision-collapse* [20], a phenomenon where training on the *base* classes force the network to discard information irrelevant for the discrimination of *base* classes, but crucial for the

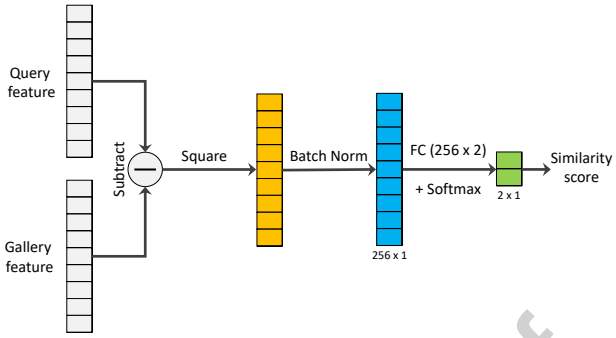


Figure 5: Query-guided Similarity Network (QSimNet) estimates the similarity score between query and gallery features. For few-shot case, these features correspond to the output of CNN in upper and lower branches (Fig. 1), for person search, they correspond to the object features generated by ID Net (Fig. 2). QSimNet is trained end-to-end with other parts of the network.

novel classes. Various *pretext* tasks have been proposed in literature for self-supervision. In this work, we opt rotation prediction [20] mainly because of its simplicity and effectiveness [20, 8]. In more details, each image in the batch is rotated by four angles (0°, 90°, 180°, 270°) and a 4-way rotation classifier is added on the top. The network is optimized with an additional rotation loss (L_{rot}), together with L_{oim} and L_{sim}. The overall loss function L_{fs} is therefore:

$$L_{fs} = L_{oim} + L_{sim} + L_{rot} \tag{5}$$

Note that we do not follow an episode based training and use the same trained model, both for the 1- and 5-shot case. The inference architecture of the 1-shot case looks similar to the training phase (without the loss functions) as shown in Fig. 1. We simply pass the query and gallery to the network to obtain their similarity score. However, in the 5-shot case, each of the 5 queries are passed to the CNN together with the gallery. This results in 5 different sets of feature vectors for each query and gallery. We compute the sum of these 5 features which are then normalised and passed to QSimNet to get the similarity score:

$$\begin{aligned} sim\_score = QSimNet(sum(f^{q_1}, f^{q_2} ... f^{q_N}), \\ sum(f^{g_1}, f^{g_2} ... f^{g_N})) \end{aligned} \tag{6}$$

where f^{g_i} is the *ith* gallery feature and f^{q_i} is the corresponding *query* (support) feature, $i = 1...N$.

3.4.2. Person search

The QGN end-to-end network training includes the detection network and the identification network, as well as QSSE, QRPN and QSimNet. The overall loss function L_{ps} is:

$$L_{ps} = L_{cls} + L_{reg} + L_{rpn_o} + L_{rpn_r} + L_{oim} + L_{qrpn} + L_{sim} \tag{7}$$

where L_{cls}, L_{reg}, L_{rpn_r} and L_{rpn_o} are the standard Faster-RCNN losses [31] for classification, regression, RPN regression and RPN objectness. The ID feature learning is supervised by standard OIM loss [3], while our new components QRPN and QSimNet are supervised by L_{qrpn} and L_{sim} respectively. The losses are shown in Fig. 2 as dark gray or dark orange boxes.

During inference, it is typical for object detection pipelines to apply NMS at the end using final classification scores. However, we use the final similarity score from QSimNet for such NMS stage during inference. The classification score from ClsNet is only used to remove least confident detections with score less than 0.01.

QRPN Anchor Sampling: Since a typical gallery image can only contain one target-person matching the query crop, the number of positive anchors is significantly fewer as compared to the negatives. This leads to a skewed positive-to-negative ratio for training of the qrpn loss (L_{qrpn}). Therefore, we augment the target person in gallery via jittering i.e. the target box is moved randomly in the nearby region. Additionally, we keep a lower anchor-to-target IoU threshold of 0.6 for positive anchor sampling. To further reduce the number of negatives, we use a batch size of 128 instead of standard 256 hence improving the positive-to-negative ratio. Note that the negative anchors are sampled from the background that do not cover other people in the gallery. This is because the non-target people in the gallery are positives for the standard RPN and it would lead to contrasting objectives for QRPN and RPN.

Table 1: Description of the five few-shot fine-grained datasets. Each row shows total number of images, total number of classes, followed by number of classes in train, val and test sets.

Dataset	**#images**	**#classes**	**#train**	**#val**	**#test**
CUB (Birds)	11,788	200	100	50	50
FGVC-Aircraft	10,000	100	50	25	25
Stanford Dogs	20,580	120	60	30	30
Oxford Flowers	8,189	102	51	25	26
Stanford Cars	16,185	196	98	49	49

4. Experimental evaluation

We experimentally evaluate QGN on recent datasets for few few-shot fine-grained classification and person search. On the few-shot fine-grained classification, QGN outperforms the current state of the art by a large margin. On the person search, QGN performs competitive with other approaches. In both cases, we provide novel qualitative visualizations of the query guidance.

4.1. Experiments on few-shot fine-grained classification

We evaluate QGN on the widely adopted Caltech-UCSD birds dataset (CUB) [10] and four other fine-grained datasets from different domains: Stanford Cars [11], FGVC-Aircraft [12], Stanford Dogs [13], and Oxford Flowers [14]. Further to evaluating various backbones, we also provide a visualization of the QSSE.

4.1.1. Benchmarks and Implementation details

The few-shot fine-grained datasets: CUB [10], Stanford Cars [11], FGVC-Aircraft [12], Stanford Dogs [13] and Oxford Flowers [14], are composed of 100-200 classes and a few thousands of images for each class. For CUB, we follow the split of [9] as used by most previous approaches. For other four datasets, we follow the split of [20]. In Table 1, we provide details of these datasets.

Evaluation Criteria: Following [8], we adopt an episodic few-shot evaluation and report the mean classification accuracy of $|D_{novel}| = 600$ randomly generated 5-way 1-shot and 5-way 5-shot episodes with $L = 15$ gallery per class.

Implementation Details: We integrate the QSSE and QSimNet modules [16] and the OIM loss [3] with the *Rotation* self-supervision of [8]. We experiment

Table 2: Comparison on the few-shot fine-grained classification task on the **CUB** dataset using 5-way. Methods below the horizontal line use either semi-supervised approach (additional unlabeled samples are used) or transductive inference (all unlabeled query samples are processed together). Our approach uses inductive inference where each query is processed independently. † denotes that the values are reported from the implementation in [9].

Setting	**Model**	**Backbone**	**1-shot**	**5-shot**	**Publication**
	MatchingNet† [23]	ResNet18	73.49	84.45	NIPS16
	MAML† [19]	ResNet18	68.42	83.47	ICML17
	ProtoNet† [7]	ResNet18	72.99	86.64	NIPS17
	RelationNet† [17]	ResNet18	68.58	84.05	CVPR18
	Baseline++ [9]	ResNet18	67.02	83.58	ICLR19
In.	S2M2 [8]	ResNet18	71.81	86.22	WACV20
	Proto+Jig [20]	ResNet18	-	89.8	ECCV20
	Baseline++ [8]	WRN	70.40	82.92	WACV20
	S2M2 [8]	WRN	80.68	90.85	WACV20
	QGN (Ours)	ResNet10	80.83	89.39	Proposed
	QGN (Ours)	ResNet18	83.82	91.22	Proposed
	QGN (Ours)	WRN	**84.15**	**91.86**	Proposed
Tran./Semi	TEAM [37]	ResNet18	80.16	87.17	ICCV19
	ICI [2]	WRN	91.11	92.98	CVPR20

with three network architectures: ResNet10, ResNet18 and WRN-28-10 (width 28, scale factor 10). Following [9, 8], the image size is 224×224 for ResNet10/18 and 80×80 for WRN. The feature embedding is 512 for ResNet10/18 and it is 640 for WRN-28-10. In all experiments, the batch size is 8 (8 query-support pairs). The negative-to-positive ratio is 3 to 1, (3 query-support samples from the same class and 1 from different ones). We train for 120 epochs using the Adam optimizer with an initial learning rate of 0.001. During training, we augment the data via random crop, image jittering and random horizontal flip.

4.1.2. Comparison to the state of the art

In Table 2, we compare QGN to state-of-the-art few-shot fine-grained classification methods on the CUB dataset. QGN with the ResNet18 backbone achieves an accuracy of 83.82 and 91.22 for the 1-shot and 5-shot cases respectively, surpassing the previous best technique S2M2 [8] by the large margins of

Table 3: Comparison on few-shot fine-grained classification on 5-way 5-shot. All models are built using ResNet18. † denotes the values are reported from the implementation in [20].

Model	**CUB**	**Cars**	**Aircraft**	**Dogs**	**Flowers**	**Publication**
Softmax†	81.5	87.7	89.2	77.6	91.0	
MAML† [19]	81.2	86.9	88.8	77.3	79.0	ICML17
ProtoNet† [7]	87.3	91.7	91.4	83.0	89.2	NIPS17
Proto+Jig† [20]	89.8	**92.4**	91.8	85.7	**92.2**	ECCV20
QGN (Ours)	**91.2**	91.3	**92.0**	**85.9**	89.9	Proposed

Table 4: Importance of each proposed model component, as evaluated on the **CUB** few-shot fine-grained classification dataset. The accuracy is reported as mean of 600 randomly generated episodes is reported.

	ResNet10		**ResNet18**		**WRN-28-10**	
Method	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Rotation [8]	-	-	72.40	84.83	77.61	89.32
OIM_R (*Baseline*)	77.76	87.88	80.27	89.81	81.45	90.15
+ *QSSE*	78.79	88.92	80.72	**91.30**	83.99	91.42
+ *QSimNet*	80.12	89.04	82.20	90.89	83.05	91.81
+ *QSSE* + *QSimNet* (=QGN)	**80.83**	**89.39**	**83.82**	91.22	**84.15**	**91.86**

12pp and 5pp. These results also surpass the performance of S2M2 with the larger WRN backbone, by 3.1pp and 0.4pp respectively. Similarly, QGN with the shallower ResNet10 backbone also surpasses S2M2 with the ResNet18 backbone by 9pp and 3.2pp. For completeness, we report in Table 2 all most recent techniques. Methods below the double line either use additional unlabeled data (semi-supervised) or evaluate all queries together (transductive), hence they do not make a fair comparison to our approach. However, these techniques appear complementary to the proposed query guidance and they could be integrated into QGN in future work.

In Table 3, we compare QGN to other approaches on four other few-shot fine-grained datasets in addition to birds (CUB). As shown in the table, for 3 out of 5 datasets i.e birds, aircrafts and dogs, we outperform the previous best results by 1.4pp, 0.2pp and 0.2pp respectively.

4.1.3. Ablation Studies

QGN components. We evaluate the effectiveness of query-guided components applicable to few-shot classification, QSSE and QSimNet, with ablation studies.

CUB. In Table 4, we consider analysis of QGN with backbones ResNet10, ResNet18 and WRN-28-10. The reference baseline combines the OIM classifier with an auxiliary rotation prediction for self-supervision. We dub this model OIM_R. This coincides with [8], which we indicate as *Rotation*, except for replacing the cosine classifier with OIM. For ResNet18, OIM_R achieves 80.27 and 89.81 for 1- and 5-shot classifications, outperforming *Rotation*, which only achieves 72.40 and 84.83. Since OIM is the leading technique for person search, but it had not been adopted for few-shot classification, this result motivates the QGN proposition for a unified approach to both tasks.

Next, we add our proposed QSSE on top of this baseline. For ResNet10, the addition of QSSE brings an improvement of almost 1pp for both 1-shot and 5-shot. For ResNet18, it brings an improvement of 0.5pp for the 1-shot and of 1.5pp for the 5-shot case. Then we add QSimNet on top of OIM_R. For ResNet10, it improves by almost 2.4pp and 1.2pp for the 1-shot and 5-shot respectively. For ResNet18, it improves by almost 2pp and 1pp. QGN for few-shot fine grained classification is given by combining QSSE and QSimNet. For ResNet10, QGN achieves an accuracy of 80.83 and 89.39, for the 1-shot and 5-shot case respectively. For ResNet18, QGN achieves an accuracy of 83.82 and 91.22. A similar improvement can be seen for the deeper WRN. Overall, in most cases, the best performance is consistently achieved by combining the two components, showing that QSSE and QSimNet are complementary.

QSSE Analysis. In Table 5, we compare the parameter and computational speed of OIM_R and $OIM_R + QSSE$. The comparison shows that the inclusion of QSSE adds only marginal additional parameters $\sim 2\%$, however runtime complexity has increased by $\sim 50\%$. This is due to the siamese design of QSSE architecture that processes pair of images together.

Table 5: Comparison of the number of parameters and runtime complexity between OIM_R and OIM_R + *QSSE*. The TFLOPS have been measured on a Tesla K80 GPU.

	Params (M)	Runtime Complexity (TFLOPS)
OIM_R	11.28	229.91
OIM_R + QSSE	11.54	344.65

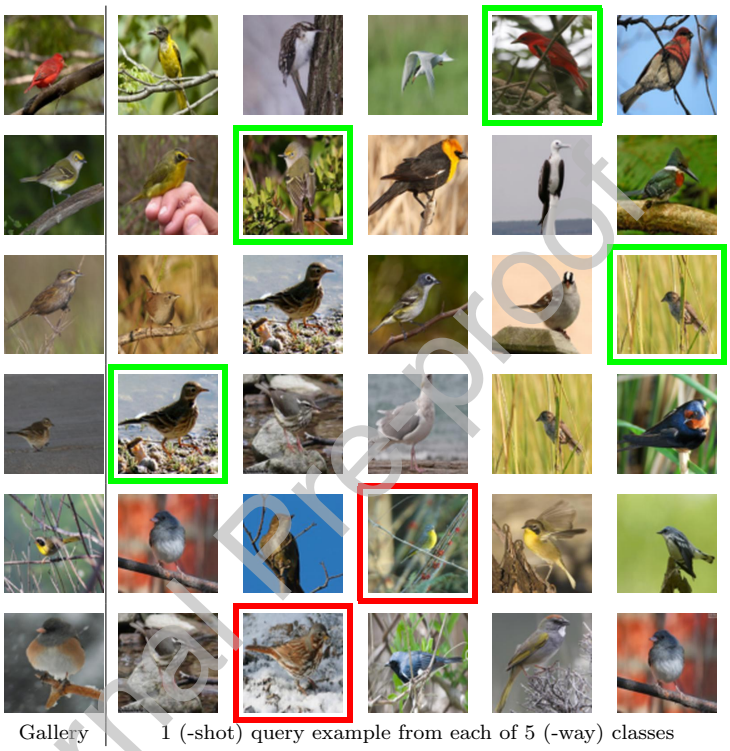


Figure 6: Qualitative results on **CUB** for 5-way 1-shot classification using our proposed QGN. The first column shows the gallery image to be classified. The next five columns show 1 (-shot) query example from each of the 5 (-way) classes. For each gallery image, the query example with highest similarity score is marked. The correctly assigned class is marked with a green bounding box, while a red bounding box depicts wrong classification.

4.1.4. Qualitative Results

In Figure 6, we illustrate some sample results of QGN for the 5-way 1-shot case on the CUB dataset. Given a gallery in the first column, we show 5 query examples from each of the 5 (-way) classes in the next 5 columns. In the first four rows, some challenging examples are given where QGN correctly classifies (green box) the gallery image. In the last two rows, there are examples where

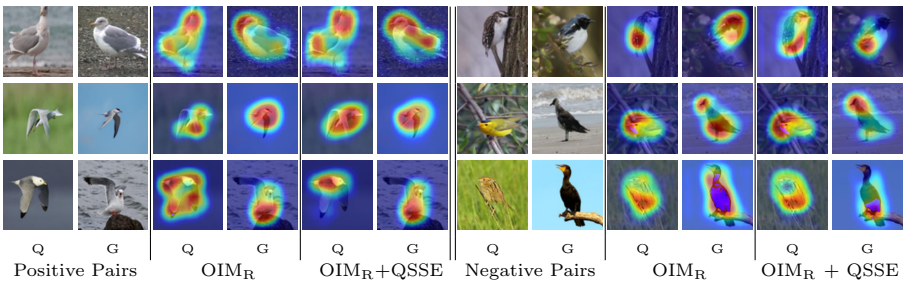


Figure 7: Class activation maps of OIM_R and OIM_R + QSSE using GradCam [38]. The left panel shows positive pairs of query (Q) and gallery (G) images from the same class; the right panel shows negative pairs. Red denotes a higher activation value while blue denotes lower. In most cases, both OIM_R and OIM_R + QSSE identify which image part to focus on (*red*-er), but OIM_R + QSSE activations are in general more accurate.

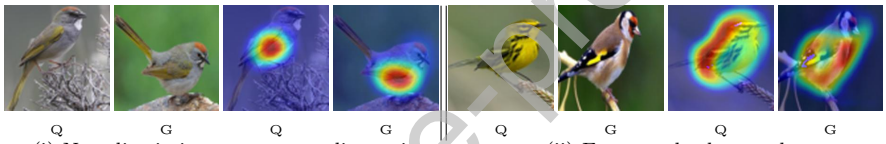


Figure 8: Class activation maps of some failure cases, where OIM_R + QSSE could not recognize the correct bird class. Q stands for query and G stands for gallery. Red denotes higher activation value while blue denotes lower. See Sec. 4.1.4 for a discussion.

QGN assigns the gallery to wrong classes (red box). Note that failure cases are also challenging for human observers, as they mainly correspond to matching front to back views of the birds.

Next, in Figure 7, we delve into the QSSE component. Using GradCam [38], we visualize some class activation maps for the OIM_R and OIM_R+QSSE models. With reference to the left panel, reporting positive query (Q) and gallery (G) pairs, note how the OIM_R+QSSE model focuses on corresponding body regions that are mostly discriminative. For example, in the first row / left panel, OIM_R+QSSE looks at the discriminant grey wing and yellow beak of the bird in both query and gallery, while OIM_R fails to focus on the wings. In the third row / left panel, high activations spread over the query example for OIM_R, while for OIM_R+QSSE high activations appear on a region which looks similar to the gallery. With reference to the right panel, reporting negative pairs, note that

the head part of the query (yellow bird) is blue in color, while that of the gallery is black, and that OIM_R+QSSE focuses only on the discriminant head part.

In Figure 8, we demonstrate some examples where OIM_R+QSSE could not recognize the correct bird. The failure happens mainly for two reasons: **i.** when the corresponding pairs attended by QSSE are not discriminative enough; and **ii.** when OIM_R+QSSE focuses on background. In general, the proposed OIM_R+QSSE finds the correct discriminative corresponding parts, better than when not using QSSE.

4.2. Experiments on Person Search

Here we evaluate QGN on the CUHK-SYSU [16] and PRW [15] datasets; we analyse quantitatively the influence of backbone architectures, input image sizes and the ROI-Pool Vs. -Align; and we illustrate the effect of QRPN.

4.2.1. Benchmarks and Implementation details

CUHK-SYSU [3] is the most used dataset for evaluating person search. It comprises 18,184 images annotated with 96,143 person bounding boxes of 8,432 identities. The training set contains 11,206 images of 5,532 identities. The test set consists of 6,978 images and 2900 queries.

PRW [15] is a dataset acquired by 6 stationary cameras in a university campus. The dataset comprises 11,816 images annotated with 43,110 bounding boxes. The training set includes 5,134 images with 482 identities, while the test set has 6,112 images with 450 identities and 2057 queries.

Evaluation metrics: Following previous works [3], we report the performance using two metrics: Common Matching Characteristic (CMC top-K) and mean Average Precision (mAP). CMC top-K is measured as the probability of retrieving at least one match in top-K predictions. Average Precision (AP) is measured for each query by calculating the area under precision-recall curve. mAP is then calculated using the mean of APs for all queries.

Implementation Details: We use OIM [3] as baseline and extend it with the three proposed query-guided components. The images are re-scaled such

that their shorter side is 600 pixels, unless mentioned explicitly. All models are trained using SGD for 4 epochs over pre-trained OIM model. The learning rate is set to 0.001, then dropped by a factor of 10 after 2 epochs. CUHK-SYSU considers as query-gallery pairs all combinations for each ID. The training set is further augmented by flipping both query and gallery images. For PRW, we sample only three gallery images for each possible query image of an ID, since the number of boxes per ID are already very large.

4.2.2. Comparison to the state of the art

In Table 6, we compare QGN to the state-of-the-art. In the top section, we report joint end-to-end methods, in the bottom section we list cascaded approaches. In each section the approaches are chronologically ordered.

As visible from the table regarding the CUHK-SYSU dataset, QGN achieves an accuracy of 91.5 mAP and 92.1 top-1, surpassing APNet [39] by 2.6pp mAP and 2.8pp top-1, BINet [33] by 1.5pp mAP and 1.4pp top-1. Following recent approaches [32, 40], we further report the performance of QGN leveraging the better FPN [41] backbone. As shown in the last row of the table, FPN+QGN achieves an accuracy of 93.7 mAP and 94.4 top-1, surpassing the most recent joint approaches including DMRNet [40] by 0.5pp mAP and 0.2pp top-1, DKD [42] by 0.6pp mAP and 0.2pp top-1. Note that FPN+QGN also performs competitive with AlignPS [32], only 0.3pp away in terms of mAP.

On PRW, QGN achieves an accuracy of 42.9 mAP and 81.9 top-1, surpassing APNet by 1pp mAP and .5pp top-1, NAE by .8pp top-1. Adopting the better FPN backbone further improves the performance. Particularly, FPN+QGN achieves an accuracy of 46.7 mAP and 82.9 top-1, surpassing NAE by 2.7pp mAP and 1.8pp top-1, PGA by 2.5pp mAP, AlignPS by 0.6pp mAP and 0.8pp top-1. Also note that FPN+QGN performs competitive to DMRNet.

4.2.3. Ablation Studies

First we evaluate the impact of QGN components, then the effect of model hyper-parameters on both OIM and QGN.

Table 6: Comparison with the state-of-the-art on the CUHK-SYSU and PRW datasets. For CUHK-SYSU, gallery size of 100 is used and for PRW the whole test set is used. Methods in the top section are joint models (*Joint*), those in the bottom are cascaded approaches (*Seq*).

		CUHK		PRW		
	Method	**mAP**	**top-1**	**mAP**	**top-1**	**Publication**
Joint	OIM [3]	75.5	78.7	21.3	49.9	CVPR17
	Context [43]	84.1	86.5	33.4	73.6	CVPR19
	APNet [39]	88.9	89.3	41.9	81.4	CVPR20
	BINet [33]	90.0	90.7	45.3	81.7	CVPR20
	NAE [5]	92.1	92.9	44.0	81.1	CVPR20
	PGA [1]	92.3	**94.7**	44.2	85.2	CVPR21
	FPN + AlignPS [32]	**94.0**	94.5	46.1	82.1	CVPR21
	FPN + DMRNet [40]	93.2	94.2	46.9	83.3	AAAI21
	DKD [42]	93.1	94.2	**50.5**	**87.1**	AAAI21
	QGN	91.5	92.1	42.9	81.9	Proposed
	FPN + QGN	93.7	94.4	46.7	82.9	Proposed
Seq.	FPN+RDLR [28]	93.0	94.2	42.9	70.2	ICCV19
	IGPN [6]	90.3	91.4	**47.2**	87.0	CVPR20
	TCTS [35]	**93.9**	**95.1**	46.9	**87.5**	CVPR20

QGN components. In Table 7, we quantify the improvements of the QGN components when integrated into OIM [3], considering two network architectures (ResNet50, ResNet18) and gallery size 100. We re-implement OIM, named *Baseline* in the table, yielding slightly better performance than [3]. As illustrated, each QGN component improves over OIM. Also, improvements are consistent for each component across different backbone architectures. Taking the representative case of ResNet50, the baseline OIM (77.2 mAP) is improved by 2.9pp with QSSE (80.1 mAP), it is improved by 2.4pp with QRPN (79.6 mAP), and by 5.4pp with QSimNet (82.6 mAP), which is the strongest single component.

QGN components are also complementary. In Table 7, considering ResNet50, QSSE+QRPN gives 82.4 mAP, QSSE+QSimNet gives 83.3 mAP, QRPN+QSimNet gives 83.1 mAP, and the full QGN set (QSSE+QRPN+QSimNet) reaches 84.4 mAP. This means an overall improvement *wrt* the baseline OIM of 7.2pp.

Reduction Ratio r of QRPN. For QRPN we choose reduction ratio r to be 16 as in [36]. Our experiments (cf. Table 8) also confirm this to be a reasonable choice as it maintains a good balance between mAP and parameter size.

Table 7: Evaluation of our proposed query-guided components on CUHK-SYSU [3] dataset. We present results for gallery size 100 using Resnet50 and Resnet18 architectures. All models in this table use an image size of 600. The OIM [3] results in the first row are taken from the original paper. OIM in the second row is our own implementation. In the last row, we report the results for our final model, OIM + QSSE + QRPN + QSimNet, which we dub as QGN.

	ResNet50		**ResNet18**	
Model	**mAP**	**top-1**	**mAP**	**top-1**
OIM [3]	75.5	78.7	-	-
OIM (Baseline)	77.2	77.6	70.0	69.7
+ *QSSE*	80.1	80.6	73.7	73.9
+ *QRPN*	79.6	80.4	73.9	73.5
+ *QSimNet*	82.6	83.0	75.3	75.3
+ *QSSE* + *QRPN*	82.4	82.8	74.7	74.4
+ *QSSE* + *QSimNet*	83.3	83.4	76.1	75.9
+ *QRPN* + *QSimNet*	83.1	83.3	75.9	75.5
+ *QSSE* + *QRPN* + *QSimNet* (= QGN)	**84.4**	**84.4**	**78.4**	**77.7**

Table 8: Person search accuracy and parameter size of OIM+QRPN ResNet18 model at different reduction ratios. We evaluate on CUHK-SYSU dataset using gallery size 100.

Ratio r	**mAP**	**top-1**	**Params (M)**
2	74.0	73.7	15.3
4	73.7	73.6	14.5
8	73.9	73.6	14.1
16	73.9	73.5	13.9
32	72.9	72.6	13.8

Hyper-parameters of OIM and QGN. In Table 9, we evaluate different design choices for OIM and QGN using the ResNet50 backbone.

CUHK-SYSU: As shown in the first few rows, the OIM baseline (77.2 mAP) improves by 3.6pp (80.8 mAP) when adopting the larger ROI pooling size 14×14 (Vs. the standard 7×7). It further improves slightly by 0.4pp (81.2 mAP) when switching to the more complex pooling method, ROI-Align. It improves by 2.7pp (83.9 mAP) when considering larger input images (smaller size re-scaled to 900 Vs. 600). Also, a larger batch size gives additional improvement taking the accuracy to 86.1 mAP (row 5). Following NAE [1], OIM may be further

[1] `https://github.com/DeanChan/NAE4PS`

Table 9: Person search accuracy on CUHK-SYSU and PRW datasets, using different design choices. For CUHK-SYSU, the standard gallery size of 100 is used and for PRW the whole test set is used. Second column gives the image size. $Pool(n)$ refers to ROI pool operation and $Align(n)$ refers to ROI align operation with output size $n \times n$. gCat refers to the concatenation of globally pooled ROI align feature with ClsIdenNet output feature (Fig. 2).

					CUHK		PRW	
Model	**imSize**	**ROI**	**bSize**	**gCat**	**mAP**	**top-1**	**mAP**	**top-1**
OIM	600	$Pool(7)$	1		77.2	77.6	29.2	65.0
OIM	600	$Pool(14)$	1		80.8	80.9	32.8	71.3
OIM	600	$Align(14)$	1		81.2	81.7	33.6	71.4
OIM	900	$Align(14)$	1		83.9	84.2	36.9	75.7
OIM	900	$Align(14)$	2		86.1	87.8	38.7	78.4
OIM	900	$Align(14)$	2	✓	88.6	88.8	40.4	79.2
QGN	900	$Align(14)$	2	✓	**91.5**	**92.1**	**42.9**	**80.9**

improved by concatenating globally pooled 1024-d features after ROI align with 2048-d feature from ClsIdenNet, bringing the OIM accuracy to 88.6 mAP. We treat this particular OIM as our baseline. Adding QGN components on top of this baseline gives our proposed model QGN, with a performance of 91.5 mAP.

PRW: Similarly on PRW dataset, largest improvements are due to increasing the pool size (32.8 Vs. 29.2 mAP), image size (36.9 Vs. 33.6 mAP), batch size (38.7 Vs. 36.9 mAP) and using finer features with gCat (40.4 Vs. 38.7 mAP). As shown in the last row, our proposed QGN gives an accuracy of 42.0 mAP.

Discussion on Runtime. Our method jointly processes each query-gallery pair. This means, for a test set of M queries and N galleries, an exhaustive search of $M \times N$ combinations is required, which makes it computationally expensive. However, note that in practical person search scenarios M is usually a small number (typically 1, i.e only one query person is being searched).

4.2.4. Qualitative results

First we compare the standard RPN [31] Vs. the proposed QRPN, then we compare OIM and QGN results.

RPN Vs. QRPN Proposals. Fig. 9 illustrates region proposals by the RPN Vs. the proposed QRPN. Given a query-gallery image pair, in column *(a)* we

(a) Query (b) RPN (c) QRPN (a) Query (b) RPN (c) QRPN

Figure 9: ***Top-10* region proposals** given by RPN and QRPN. Ground-truth boxes are in yellow, output region proposals are in blue. (a) Query images with the queried person ground-truth box, (b) Gallery images with RPN proposals (c) Gallery images with QRPN proposals.

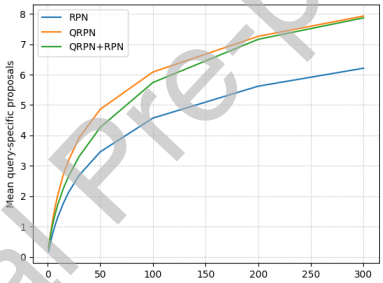


Figure 10: Average query-specific proposals in top-N proposals for the RPN, QRPN and RPN+QRPN sub-networks.

show the query images with the person bounding boxes (in yellow). In columns *(b)* and *(c)* we illustrate the top 10 region proposals in the gallery by RPN and QRPN, respectively. Note that the proposals by the RPN are on any person in the image, as it is trained for generic person detection. By contrast, the QRPN proposals in column *(c)* are query-guided and are focused on those people which mostly resemble the queried person. Specific examples are the second row/left panel and the third row/right panel, where QRPN specifically proposes people wearing clothes of the same color, and the last row/right panel where RPN fails due to contrast challenges while QRPN leverages the query person pattern and

Figure 11: Qualitative *Top-1* person search results for a number of challenging examples. For each example, we show (a) the query images with the bounding box of the query-person, in yellow, (b) their corresponding output matches given by the baseline OIM, and (c) results of our proposed QGN. Red bounding boxes are failures, green ones represent correct matches.

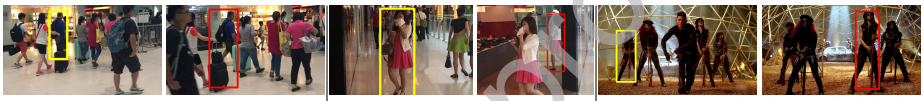


Figure 12: Typical failures: (a) localization error, (b) missing annotation, and (c) a challenging example with similarly-looking people.

successfully estimates regions over it.

We support the qualitative result with Fig. 10, i.e. a plot of the number of query-specific proposals (y-axis) among the top-N proposals (x-axis). A query-specific proposal is one that has IoU >= 0.5 with the target, one which serves to detect the queried person. Note how QRPN and QRPN+RPN consistently provide more query-specific proposals than the standard RPN. Additionally, training with both QRPN and RPN sub-networks results in better performances.

OIM Vs. QGN. Fig. 11 illustrates some challenging queries (column *(a)*) and gallery images, where these are searched for, either with OIM (column *(b)*) or QGN (column *(c)*). Top-1 search results are reported. Note how QGN retrieves a query person from a crowd (first row / left panel), distinguishes a query person from similarly dressed ones (second row / right panel), and also re-identifies the query in low contrast and illumination conditions (third row / right panel).

In Fig. 12, we illustrate typical failure cases of QGN. In *(a)*, QGN successfully retrieves the correct person, but the bounding box is poorly aligned (IoU < 0.5).

(b) is an interesting case of missing annotation for the target person, i.e. QGN detects the reflection of the girl in the mirror, which is considered false positive. *(c)* is challenging due to the similar appearance and low visibility of the people.

5. Conclusion and Future Work

This work has addressed, for the first time, few-shot fine-grained classification and person search with a unified Query-Guided Network (QGN). Uniting best practices from the two tasks has allowed QGN to define a novel state-of-the-art in few-shot fine-grained classification and to be on par with it for person search. A second contribution has been to propose query guidance via three components, which may be plugged-in at various stages of classification and detection models. Query guidance is novel for few-shot fine-grained classification, and it has been shown effective both quantitatively and qualitatively. In person search, query-guidance had been the novel introduction of our work [16], now adopted by various state-of-the-art techniques, which we re-state here as effective. A drawback of our approach is its computational complexity which is due to the interaction of a pair of images at all levels in the network, notably in the Siamese QSSE network. In future work, following the spirit of a unified query-guided framework, we plan to research few-shot fine-grained detection, for which the query-guided proposal network module of QGN may also be relevant.

6. Acknowledgments

This work is partially supported by Sapienza (Bandi d'Ateneo) and by the project of the Italian Ministry of Education, Universities and Research (MIUR) "Dipartimenti di Eccellenza 2018-2022".

References

[1] H. Kim, S. Joung, I.-J. Kim, K. Sohn, Prototype-guided saliency feature learning for person search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4865–4874.

[2] Y. Wang, C. Xu, C. Liu, L. Zhang, Y. Fu, Instance credibility inference for few-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12836–12845.

[3] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, Joint detection and identification feature learning for person search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3415–3424.

[4] D. W. L. Tang, B. Hariharan, Revisiting pose-normalization for fine-grained few-shot recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14352–14361.

[5] D. Chen, S. Zhang, J. Yang, B. Schiele, Norm-aware embedding for efficient person search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12615–12624.

[6] W. Dong, Z. Zhang, C. Song, T. Tan, Instance guided proposal network for person search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2585–2594.

[7] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 4077–4087.

[8] P. Mangla, N. Kumari, A. Sinha, M. Singh, B. Krishnamurthy, V. N. Balasubramanian, Charting the right manifold: Manifold mixup for few-shot learning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2218–2227.

[9] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. Wang, J.-B. Huang, A closer look at few-shot classification, in: International Conference on Learning Representations (ICLR), URL `https://openreview.net/forum?id=HkxLXnAcFQ`, 2019.

[10] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200-2011 dataset, Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011).

[11] J. Krause, M. Stark, J. Deng, L. Fei-Fei, 3d object representations for fine-grained categorization, in: IEEE International Conference on Computer Vision Workshops (ICCVW), 2013, pp. 554–561.

[12] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, A. Vedaldi, Fine-grained visual classification of aircraft, Tech. rep. (2013). arXiv:1306.5151.

[13] A. Khosla, N. Jayadevaprakash, B. Yao, L. Fei-Fei, Novel dataset for fine-grained image categorization, in: Workshop on Fine-Grained Visual Categorization (FGVC), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[14] M. E. Nilsback, A. Zisserman, A visual vocabulary for flower classification, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2006, pp. 1447–1454.

[15] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, Person re-identification in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1367–1376.

[16] B. Munjal, S. Amin, F. Tombari, F. Galasso, Query-guided end-to-end person search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 811–820.

[17] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, T. M. Hospedales, Learning to compare: Relation network for few-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1199–1208.

[18] Z. Zhao, Q. Liu, W. Cao, D. Lian, Z. He, Self-guided information for few-shot classification, Pattern Recognition 131 (2022) 108880.

[19] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: Proceedings of the 34th International Conference on Machine Learning (ICML), 2017, pp. 1126–1135.

[20] J.-C. Su, S. Maji, B. Hariharan, When does self-supervision improve few-shot learning?, in: Proceedings of the European Conference on Computer Vision (ECCV), 2020, pp. 645–666.

[21] R. Hou, H. Chang, B. Ma, S. Shan, X. Chen, Cross attention network for few-shot classification, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019, pp. 4005–4016.

[22] S. Y. F. Zhao, J. Zhao, J. Feng, Dynamic conditional networks for few-shot learning, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 19–35.

[23] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, D. Wierstra., Matching networks for one shot learning, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016, pp. 3637–3645.

[24] M. Liang, S. Huang, S. Pan, M. Gong, W. Liu, Learning multi-level weight-centric features for few-shot learning, Pattern Recognition 128 (2022) 108662.

[25] Y. Zhu, C. Liu, S. Jiang, Multi-attention meta learning for few-shot fine-grained image recognition, in: Proceedings of the International Joint Conference On Artificial Intelligence (IJCAI), 2020, pp. 1090–1096.

[26] H. Tang, C. Yuan, Z. Li, J. Tang, Learning attention-guided pyramidal features for few-shot fine-grained recognition, Pattern Recognition 130 (2022) 108792.

[27] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, Y. Wei, Circle loss: A unified perspective of pair similarity optimization, in: Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6398–6407.

[28] C. Han, J. Ye, Y. Zhong, X. Tan, C. Zhang, C. Gao, N. Sang, Re-id driven localization refinement for person search, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9814–9823.

[29] B. Munjal, F. Galasso, S. Amin, Knowledge distillation for end-to-end person search, in: Proceedings of the British Machine Vision Conference (BMVC), BMVA Press, 2019, pp. 31.1–31.16.

[30] C. Liu, H. Yang, Q. Zhou, S. Zheng, Making person search enjoy the merits of person re-identification, Pattern Recognition 127 (2022) 108654.

[31] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Proceedings of the 29th International Conference on Neural Information Processing Systems, 2015, pp. 91–99.

[32] Y. Yan, J. Li, J. Qin, S. Bai, S. Liao, L. Liu, F. Zhu, L. Shao, Anchor-free person search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7690–7699.

[33] W. Dong, Z. Zhang, C. Song, T. Tan, Bi-directional interaction network for person search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2839–2848.

[34] W. Li, S. Gong, X. Zhu, Hierarchical distillation learning for scalable person search, Pattern Recognition 114 (2021) 107862.

[35] C. Wang, B. Ma, H. Chang, S. Shan, X. Chen, Tcts: A task-consistent two-stage framework for person search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11952–11961.

[36] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141.

[37] L. Qiao, Y. Shi, J. Li, Y. Wang, T. Huang, Y. Tian, Transductive episodic-wise adaptive metric for few-shot learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3603–3612.

[38] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.

[39] Y. Zhong, X. Wang, S. Zhang, Robust partial matching for person search in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6826–6834.

[40] C. Han, Z. Zheng, C. Gao, N. Sang, Y. Yang, Decoupled and memory-reinforced networks: Towards effective feature learning for one-step person search, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 1505–1512.

[41] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2117–2125.

[42] X. Zhang, X. Wang, J. Bian, C. Shen, M. You, Diverse knowledge distillation for end-to-end person search, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 3412–3420.

[43] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, X. Yang, Learning context graph for person search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2158–2167.

Declaration of interests

☐The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

tum.de, in.tum.de, osram.com, di.uniroma1.it, uniroma1.it, google.com, ethz.ch

Biographies

Bharti Munjal is a PhD student at the Technical University of Munich (TUM), Germany. She received her engineering degree from the National Institute of Technology Kurukshetra India and her masters degree from Technical University of Munich. Her research interests include object detection, person search, person re-identification, few-shot and meta learning.

Alessandro Flaborea is a PhD student at the Sapienza University of Rome. He earned his degree in Computer Science from the University of Udine, Italy and received his master degree in Data Science from the Sapienza University of Rome. His research interests include few-shot learning, anomaly detection and geometric deep learning.

Sikandar Amin is leading R&D efforts in computer vision and deep learning at Magic Leap, Inc. His interests include object detection and tracking, few-shot fine-grained classification, re-identification and full body pose estimation. He obtained his PhD from Technical University of Munich in 2D and 3D human pose estimation.

Federico Tombari is a Research Scientist and Manager at Google and a Lecturer at Technical University of Munich. He has 200+ peer-reviewed publications in the field of computer vision and machine learning. He has received two Google Faculty Research Awards, one Amazon Research Award, two CVPR Outstanding Reviewer Awards.

Fabio Galasso heads the Perception and Intelligence Laboratory at the Sapienza University of Rome. He is interested in research and innovation transfer in perception, multi-agent systems and general intelligence in sustainable interpretable frameworks. Specific topics include detection, tracking, forecasting, anomaly-detection, re-identification, few-shot and meta-learning, domain adaptation and multi-modal computing.