MFSC: Matching by Few-Shot Classification

Daniel Shalam[1]
dani360@gmail.com

Elie Abboud[1]
eliabboud1000@gmail.com

Roee Litman[2]
github.com/rlit

Simon Korman[1]
skorman@cs.haifa.ac.il

[1] University of Haifa
Haifa, Israel

[2] Rafael Advanced Defense Systems
Tel-Aviv, Israel

Abstract

The ability to accurately and efficiently match between sets of items has always been fundamental in computer vision pipelines and applications with a wide variety of realizations that involve finding correspondences between sets of local features, small patches or entire images collections. In recent years, the emergence of deep learning has facilitated significant improvements of matching based applications. This progress was achieved through advancing improved data embedding and description, and less focus was put on the matching process itself. Specifically, relying on simple pairwise or triplet distance-based metric learning, ignoring the set-to-set nature of the problem.

We suggest a holistic approach to matching, by observing its natural connection to few-shot classification (FSC), a largely growing research area that deals with learning using limited amounts of data. We argue that certain popular FSC paradigms, such as meta-learning and transductive learning, are particularly suitable for tackling the specific challenges that arise in matching problems. Our approach, MFSC, builds upon state-of-the-art features and FSC algorithms, significantly improving the quality of matching. Moreover, we show how to construct a meta-learning scheme based on our approach, which allows end-to-end training of the entire matching process. We validate our method on the tasks of patch-correspondence, image-alignment and person re-identification.

1 Introduction

We deal with the problem of finding a matching or correspondence between two sets of items. Of particular interest, is the case of matching between sets of images [1, 42], local features [1, 33] or patches [4, 19]. This problem appears abundantly in computer vision applications, as a mid-level or high-level task, and therefore the development of accurate and efficient matching techniques is of high importance. Given two sets of items $\mathcal{A}$ and $\mathcal{B}$, of cardinality N and M respectively, indexed by $[N] := \{1,2,...,N\}$ and $[M] := \{1,2,...,M\}$, the goal is to recover an underlying *matching*, which is a set of *matches*, or pairs of items - one from each set. The matches are typically assumed to be mutually exclusive, meaning that each item belongs to at most one match. Such a matching can be defined by a mapping $\pi_{\mathcal{B}} : [M] \rightarrow \{0\} \cup [N]$, where item j in $\mathcal{B}$ is matched to item i in $\mathcal{A}$ if and only if $\pi_{\mathcal{B}}(j) = i$.

In computer vision, a common example is the correspondence between a pair of related images, in the form of pairs of matching interest points that were independently detected in each of the images as the basis for most geometric fitting and alignment algorithms. Another example, at the application level, is the relation between sets of images of different items, identities or classes. This happens when comparing between face corpuses, or sets of people captured by different surveillance cameras (person re-id [42]), between animal or plant image reference guides, or between sets of alphabet characters handwritten by different individuals.

We suggest a new perspective on the matching problem, by framing it as a *few-shot classification* task. The domain of few-shot learning (FSL), and few-shot classification (FSC) in particular, has enjoyed a tremendous amount of research over the last years [11, 12, 28, 31, 39, 44, 47]. A wide variety of techniques have been developed to face the difficulty in training standard supervised machine learning methods on limited amounts of data. This situation occurs when data-sets are either small, or long-tail distributed and is related to the settings of transfer-learning and domain adaptation.

Why do we think of matching in terms of FSC? Matching can be considered, a priori, to be an unsupervised task since at test time inference is not done with respect to any prior training data, but rather between data (e.g. images of objects) that has not been seen before. On the other hand, examples of matched data can be used to learn *how* to match correctly, in a supervised manner. This reasoning naturally leads us to consider the episodic-based *meta-learning* (or "learn to learn") framework, where a matcher can be trained on offline matching tasks, in a way that it can efficiently adapt to new test-time tasks.

Another paradigm extensively used in FSC, in order to fully exploit additional sources of information, is that of *transductive learning*. The use of auxiliary information in the unlabeled test data has been shown to provide a significant source for improvements in the data-scarce setting. Transductive learning techniques are especially relevant to matching, where most of the information is present in the task itself, including certain regularities that should be exploited, such as the one-to-one matching relations.

We initiate the study of image matching using few-shot learning techniques. Our main contribution is in making this possibility explicit, by a rather simple reduction, that allows employing a variety of general FSC tools and developing some dedicated ones. Our techniques are demonstrated on 3 very different settings of the problem, suggesting its wide applicability and further potential.

2 Related Work

2.1 Few-Shot Classification

Few-Shot-Classification (FSC) is a branch of few-shot-learning in which a classifier needs to learn to recognize previously unseen classes given a limited number of labeled examples. A FSC task [39] is a self-contained instance that includes both support (labeled) and query (unlabeled) items. In the *meta-learning* approach, the training data is split into tasks (or episodes) mimicking the test time tasks to which the learner is required to generalize. MAML [12] “learns to fine-tune" by learning a network initialization from which it can quickly adapt to novel classes. In ProtoNet [34], a learner is meta-trained to predict query feature classes, based on distances from support class-prototypes in the embedding space. The trainable ’meta-learning’ version of SOT can be viewed as a meta-learning algorithm.

Subsequent works [8, 11] advocate fine-tuning pre-trained networks, with larger and more expressive backbones, and employ *transductive* inference, which fully exploits the data at inference, including unlabeled images. Finally, a significant number of works [16, 17, 47] have adopted the Sinkhorn algorithm [9] as a parameterless unsupervised classifier that computes matchings between query embeddings and class centers. Laplacian-Shot [47] is based on minimizing a unary term that assigns query samples to nearest class prototypes, and a pairwise Laplacian term that encourages label consistency between nearby query samples. In PTMAP [16] the features are first preprocessed to better align to a Gaussian distribution and then undergo an optimal-transport based iterative algorithm to estimate the class centers.

2.2 Matching Applications

Matching of Local Patches The introduction of local patch matching benchmarks has encouraged rigorous, reproducible, and large scale experimentation on local feature description and matching. Benchmarks like PhotoTourism [41], RomePatches [29] and HPatches [4] are large and diverse, facilitating the growing needs of learning based methods. Patch based benchmarks consist of patches extracted from interest point locations in images that are size-normalized, and annotated pair- or group-wise with labels that indicate positive or negative examples of correspondence. They were used extensively to evaluate local descriptors, whether handcrafted [23] or learned [25, 36, 37] and matching algorithms [15, 20].

In this work we focus on the most challenging task of the HPatches benchmark [4] - the 'image matching' task, which contains millions of patches extracted from hundreds of images pairs. Each matching task consists of matching between the entire set of (thousands of) extracted feature points as they appear in a pair of related images, after the addition of synthetic geometric distortions. We improve upon nearest-neighbor matching of state-of-the-art descriptors such as HardNet++ [25] and SOSNet [37] that were trained using advanced model architectures and triplet-based deep metric learning formulations.

Pose Estimation The task of estimating the Fundamental matrix between a pair of uncalibrated images of a scene, has been a long-standing challenge, allowing to stress-test each of the stages that are needed to regress the relative pose between the cameras. This includes feature detection, description, robust matching and fitting algorithms.

While the classical pipeline of SIFT and RANSAC with the 8-point algorithm is still a strong baseline, improvements have been achieved using deep-learning detectors and descriptors such as HessAffNet [26] and HardNet++ [25], pruning methods [6, 22] and robust estimators (e.g. LMedS [32] and MLESAC [38]). FM-Bench [5] is a feature matcher evaluation benchmark that compares many relevant algorithms, over four datasets [14, 18, 35, 40].

Person Re-Identification (Re-ID) is the task identifying a certain person (identity) between multiple detected pedestrian images, from different non-overlapping cameras. It is challenging due to the scale of the problem and large variation in pose, background and illumination. Re-ID is typically considered an instance retrieval problem and hence can be tackled using metric learning tools. The data is divided into a set of query images and a large set of gallery images, with the goal of finding a representation that minimizes the relative distances between matching samples. See [43] for an excellent recent and comprehensive survey.

OSNet [46] developed an efficient small-scale network with high performance and the two-branch structure Batch DropBlock (BDB) Network [10] and its extension (Top-DB-Net) [30] learn comprehensive and spatially distributed features consisting of both global and attentive local representations. From our point of view, this task differs from the others considered in that it is larger scale (querying thousands of identities against a target of tens of thousands) and more real-world compared to the carefully curated FSC sets.

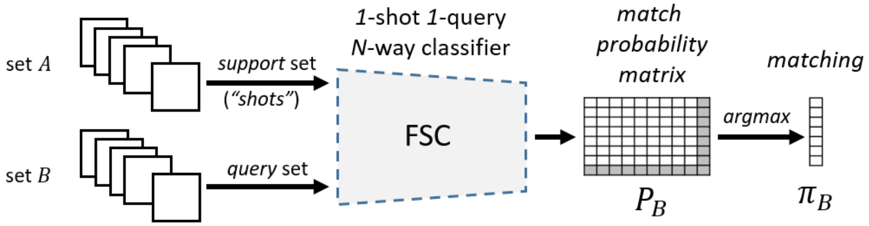

Figure 1: **Matching by Few shot Classification (FSC)**: Any few-shot classifier (FSC) can be used for matching between two sets of items. A 1-shot 1-query N-way classifier receives a support set (of labeled examples) consisting of one example of each of N classes (given by image set A). Its goal is to classify the items of an unlabeled query set with (at most) one query per class (given by items set B). Its output on the entire query set can be arranged in a match probability matrix, with one row per query, and the final matching $\pi_{\mathcal{B}}$ is obtained by taking an argmax on the matrix rows.

3 Matching by Few-Shot Classification (FSC)

A basic observation that we make, is that a typical matching problem can be framed as a few-shot classification problem. The FSC problem has been standardized [39] as follows: In the N-way K-shot Q-query classification problem, a *task* consists of a classification instance, for which the learner has access to K labeled examples from each of N classes and is tested on Q unlabeled examples per class. The problem becomes more challenging as the number of ways N increases and as the number of shots K decreases[1]. Typical settings are 5-way or 20-way in combination with 5-shots or 1-shot, with up to 15 queries.

The problem of *matching* between sets of size N and M can naturally be posed as an N-way 1-shot classification task. Consider two sets of items to be matched, A and B, where a motivating example would be sets of interest-points (e.g. patches) extracted independently from a pair of images in a two-view (stereo) setting. Each item in A can be thought of as the *single* example representative of its own class[2]. Finally, an item in B can be matched to A by *classifying* it with respect to the N classes defined by the items in A.

This concept is illustrated in Figure 1, showing how any few-shot classifier (FSC) can be used for matching between sets of items. A 1-shot 1-query N-way classifier receives a support set (of labeled examples) consisting of one example of each of N classes (given by image set A). Its goal is to classify the items of an unlabeled query set with (at most) one query per class (given by image set B). Its output on the entire query set can be arranged in a match probability matrix, with one row per query, and the final matching $\pi_{\mathcal{B}}$ is obtained by taking an argmax on the matrix rows. While any FSC method could be plugged in to this framework, we next look into the desired properties of such a classifier and suggest a specific one in the following.

3.1 Design Considerations

Posing a matching problem as an N-way 1-shot classification problem comes with the large benefit of being able to use recent FSC machinery to tackle the matching problem, with promising results that we show below.

While the reduction between the problems is seemingly simple, several unique properties

[1] Some methods exploit the test samples as well and hence might improve as the number of queries Q increases.

[2] In the two-view setting, a patch is a single example of the class of image realizations of a particular 3D scene location - Such classes are well defined, since different 2D image locations are reprojections of distinct 3D points

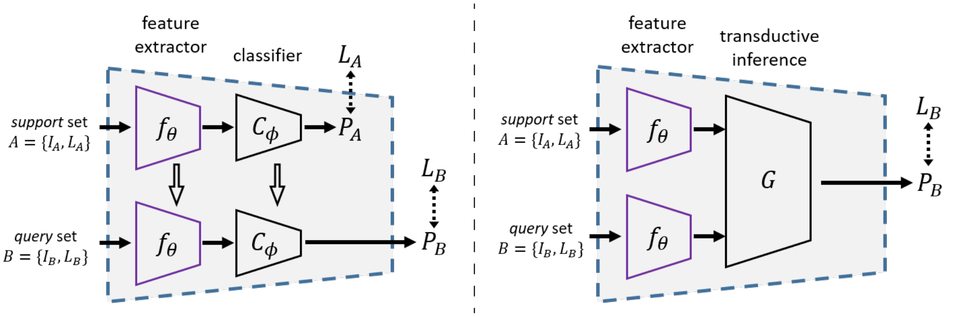

Figure 2: **Few-Shot Classification (FSC) paradigms**: **non-transductive** (left) and **transductive** (right) FSC designs differ in whether the unlabeled query set is used at *test* time. In **meta-learning**, the learning of the inference parameters (θ and ϕ) is done at training time, by using the exact same test time inference process, except that the query labels $L_{\mathcal{B}}$ are available for supervision.

of the matching problem must be taken into account. (i) *symmetry*: Matching between two sets is a symmetric problem. Therefore, learning on one set and classifying the other with respect to the first is likely sub-optimal; (ii) *injectivity*: A matching task can often be assumed to be injective (as in the case of patches from a pair of images), as opposed to the general case of classification, where different queries might belong to the same class. This is a strong constraint that should be exploited, i.e. the classification of an item is not independent from the classification of the rest; (iii) *available data at test time*: In classification, queries are typically given in an online incremental fashion and hence the target set cannot be used for fine-tuning, which this is not the case for matching; (iv) *partiality*: each query belongs to some class, while some items might not be matched.

In addition to the differences from standard FSC mentioned above, real world setups of the matching version tend to be more challenging in terms of the general setting. First, FSC is typically tested on randomly generated groupings of a *small* number of classes (e.g. up to 20) [12, 28, 31, 39, 44], while in image matching one needs to deal with a very large number of classes (e.g. interest points per image, up to thousands). Second, the number of classes might substantially vary across a collection of matching tasks (e.g. in the range of hundreds to thousands when matching between images) and can not be assumed to be constant throughout the learning process. Third, we are in the case of 1-shot and 1-query, which is the most extreme in terms of information per class at training and testing, respectively.

Transductive Learning In Transductive FSC, the idea is to use information from the query set, at test time, to guide the learning of the classifier, which is largely based on the support set. This approach is especially suitable for '1-shot-many-way' data-scarce case of matching. In Figure 2 we show generic designs of non-transductive and transductive FSCs. These designs are rather general and can be used to frame most non-transductive and transductive FSC methods (e.g. [12, 28, 34, 39, 44] and [8, 11, 16, 47] respectively).

In the non-transductive setting (Fig. 2 left), the inference pipeline which includes a feature extractor and a classifier is trained on the labeled support set and then applied to the unlabeled query set. This fits the general classification setup, where queries are not available in advance and are given in an online manner, which is not the case in matching. Furthermore, this approach is very asymmetric in nature, which is, as previously discussed, sub-optimal for the matching case. In the transductive setting (Fig. 2 right), both image sets undergo feature extraction followed by a joint processing that outputs the query classification results. This approach enables exploiting the information in the query samples, e.g. for jointly estimating class distributions or for enforcing the injectiveness of the matching.

Meta Learning By making prediction conditioned on a given (support) set of examples, meta-learning [39] methods can 'learn to learn' from limited labeled data by meta-training on a collection of labeled tasks (episodes), which are of the exact same structure of those in the meta-testing stage, where the classification model adapts to predict novel classes based on a new support set of examples. Methods of this kind are typically distance metric learning based [34, 39] or initialization based, such as the seminal MAML [12].

In our setting, we experiment with using meta-learning after reducing a matching task to a few-shot task, with the goal of obtaining better generalization between matching tasks. In particular, rather than using feature extractors that have been trained using a separate protocol on the entire training set, we consider episodic training, where in a meta-training stage random matching tasks are generated from the train set to imitate the test tasks.

Partial Matching A major challenge that one needs to consider in many practical matching tasks, is the fact that certain items in one set will not have a match in the other. This is in contrast to the typical classification setting, where each item is assumed to belong to one of the given classes and usually each class is assumed to have representative items in the data.

In the field of FSC, which we are building upon, the datasets are curated in the way-shot-query setting, which implies a full and balanced matching between items and classes. We aim to tackle the more general setting, where items from either of the sets might not be matchable. In order to do so, we will add a dummy node (or 'dustbin') item in each of the sets, which can be matched to an unlimited number of items from the other set. The output of our few-shot matcher will include assignment probabilities to these dummy nodes, as is depicted by the shaded last row and column of P_B, in Figure 1.

3.2 A (Meta-)Transductive Matching Algorithm

In this section we describe our FSC-based matching algorithm, outlined in Fig. 3. Following the arguments in the previous section, we favor a transductive approach, that follows the design in Fig. 2 right. We start by explaining the transductive FSC block (Fig. 2 right) which is referred to as procedure ***Transd-FSC*** in Algorithm **MFSC**. Notice that it can be executed directly at inference using the 'testing' mode if provided with a pretrained feature extractor f, or one can first use the 'meta-training' mode to train a feature extractor.

We present here a particular transductive FSC algorithm, which is based on the recent PTMAP [16] (while stressing that our approach is general and could be based on other transductive methods that follows the design in Fig. 2 right, such as [8, 11, 47]). The main idea, following [16] (and similarly [47]) is to iteratively estimate "class centers" for each of the classes represented by the items in set A and the assignments of items in B to these centers.

In detail, an input pair of sets of items $A = \{a_1,...,a_N\}$ and $B = \{b_1,...,b_M\}$, is embedded into a common feature-space, by a given trained feature extractor f, resulting in $f_A = \{f(a_1),...,f(a_N)\}$ and $B = \{f(b_1),...,f(b_M)\}$. These features then go through a power-transform of the form $PT(v) = (v+\varepsilon)^\beta/||(v+\varepsilon)^\beta||$ where $\beta > 0$ is a hyper-parameter that controls the skewness of the feature distribution and $\varepsilon > 0$ is a small constant for used for stability. For more details refer to [16]. The class centers C are initialized with f_A (step 3.), the resulting features of set A, followed by an iterative update of: image-to-class distances (step 4. (a)), image-to-class assignments (step 4. (b)), where the images in set B are fractionally assigned to these classes (each image is assigned a vector of probabilities that sums

```
Algorithm MFSC (matching tasks $\{(A_i, B_i)\}$; learnt/pretrained extractor $f$; labels (train only) $\{L_i\}$)
  1. $P_i$ = Transd-FSC($A_i$, $B_i$, $f$)  # inference over batch (per $i$)
  2. if 'testing' return $\arg\max(P_i)$  # matching (predictions vector, per $i$)
  3. if 'meta-training'
     (a) $\ell = \sum_i CEL(P_i, L_i)$  # calculate cross-entropy loss
     (b) update and return extractor $f$  # through back-propagation
```

```
Transd-FSC (image sets $A$, $B$; extractor $f$)
  1. $f_A = f(A)$; $f_B = f(B)$;  # feature extraction
  2. $f_A = PT(f_A)$; $f_B = PT(f_B)$;  # power transform
  3. $C = f_A$  # initialize class centers
  4. repeat $k$ times:
     (a) $D = dists(f_B, C)$  # $N \times M$ feature-center $\ell_2$ dists
     (b) $P_B$ = Sinkhorn$(D, mode)$  # opt. transport
         (remove last non-match row/col in 'partial' mode)
     (c) $C = \frac{\alpha}{2}(f_A + f_B P_B) + (1-\alpha)C$  # update centers
  5. return $P_B$  # match prob. matrix
```

```
Sinkhorn (distance matrix $D$, partiality mode $mode$)
  1. $c = \mathbf{1}_{M\times 1}$; $r = \mathbf{1}_{N\times 1}$;  # target col and row sums
  2. if $mode$=='partial':  # as opposed to 'full'
     (a) D = $\begin{bmatrix} D & \delta\mathbf{1}_{N\times 1} \\ \delta\mathbf{1}_{1\times M} & 0 \end{bmatrix}$ ;  # non-match penalties
     (b) $c = [\mathbf{1}_{1\times M}\ N]^T$; $r = [\mathbf{1}_{1\times N}\ M]^T$;  # row/col sums
  3. $K = e^{-\lambda D}$;  # $K$ matrix
  4. repeat $t$ times:  # init $u$, $v$ to all-1 vectors
         $v = c./K^T u$; $u = r./Kv$;  # normalize cols, rows
  5. return $diag(u) \cdot K \cdot diag(v)$  # transport plan
```

Figure 3: **Pseudocode of *MFSC* - Matching by FSC with sub-routines *Transd-FSC* and *Sinkhorn*.** It operates in either 'testing' (inference) or 'meta-training' (learning) mode, by matching between the sets A_i and B_i, using a feature extractor f that is either learnt or pretrained. The ***Transd-FSC*** call (step 2.) performs the few-shot inference (here implemented in the lines of the PTMAP [16] algorithm). Meta-learning is performed (step 3.) by updating the embedding f using cross-entropy loss.

to 1) and class center updates (step 4. (c)), where each item contributes, with a weight that is proportional to its assignment probability, to the re-estimation of each class center.

The fractional assignment itself (step 4. (b)) is found using the Sinkhorn algorithm [9] which approximates the optimal assignment of images to classes, under the updated pairwise distance matrix between images and classes. We provide a detailed implementation of the Sinkhorn algorithm, one which can explicitly handle both 'full' (permutation like) or 'partial' matching modes. Partial matching is achieved by padding the distances matrix D (and transport plan) by a row and column, which allow for a source or target item to be unmatched. δ is a parameter that determines the cost of not matching an item, which we set heuristically to equal the maximal value of D.

4 Results

4.1 Matching of Patches

In this section we evaluate the suggested approach by experimenting on the challenging 'image matching' task of the HPatches dataset [4], which contains over 2.5 million patches extracted from 116 image sequences, each with 6 images with known homography. Each matching task consists of matching between the entire set of extracted feature points (in the form of patches, typically up to 2000) as they appear in a pair of related images, after the addition of synthetic geometric distortions.

We have implemented the reduction from matching to FSC for three different FSC methods [12, 16, 47] over several leading feature embeddings [25, 27, 37]. The reduction is to a 1-shot-1-query-M-way instance, with target/source image patches serving as shots/queries respectively, where M is the number of patches, which is equal for source and target images in this dataset, hence using the mode='full' option in step 4b of Alg. ***Transd-FSC***. Our

Table 1: **mean average-precision (mAP) on HPatches [4] matching task**. Blocks 1-2: Handcrafted and learned descriptors, with nearest neighbors based matches. Blocks 3-5: MFSC based matching. Blocks 3 and 5 involve MFSC meta-training. **bold**/<u>underlined</u> results are best/second-best per column. The 'meta' methods involve fine-tuning of the features on the 'train' split.

method	**category**	**'easy'**	**'hard'**	**'tough'**	***mean***
SIFT [23]	handcrafted	46.5%	20.3%	9.7%	25.5%
Root-SIFT [3]		49.4%	22.0%	9.7%	27.1%
Hnet++ [25]	learned	72.2%	56.2%	37.9%	55.4%
Hnet-PS [27]		69.3%	58.6%	44.6%	57.5%
L2-Net [36]		73.0%	57.5%	39.1%	56.6%
DOAP-ST-LM [15]		74.5%	66.9%	57.0%	66.3%
SOSNet [37]		76.3%	68.4%	56.5%	67.1%
MFSC-MAML [2, 12]	meta	77.8%	65.9%	50.8%	64.9%
MFSC-LapShot [47] (HNet++)	transductive	79.9%	70.3%	55.1%	68.3%
MFSC-PTMAP [16] (HNet++)		82.2%	71.6%	52.8%	68.8%
MFSC-PTMAP [16] (HNet-PS)		79.5%	73.1%	61.6%	71.4%
MFSC-PTMAP [16] (SOSNet)		<u>84.3%</u>	<u>80.0%</u>	<u>71.6%</u>	<u>78.6%</u>
MFSC-PTMAP [16] (HNet++)	meta + transductive	77.9%	73.0%	63.9%	71.6%
MFSC-PTMAP [16] (SOSNet)		**85.2%**	**81.1%**	**73.1%**	**79.8%**

MFSC variants are compared to the baseline nearest-neighbor matching based results with the respective features, either 'learned' (trained on HPatches) or 'handcrafted'.

The results are aggregated in Table 1. We begin by using the PTMAP [16] based ***Transd-FSC*** with HardNet++ [25] or SOSNet [37] features, within Alg. ***MFSC***, with either 'testing' only (e.g. pre-trained features), denoted by 'transductive' or with the preliminary 'meta-training', denoted by 'meta+transductive'. In addition, we evaluate using the 'LapShot' version of ***Transd-FSC*** (with Laplacian-Shot [47] replacing PTMAP as the FSC algorithm), or using the improved version [2] of MAML [12], which is a non-transductive meta-learning FSC. Also, we consider applying MFSC-PTMAP on Hardnet-PS [27] (Hardnet features pre-trained on the extensive patch-matching PS dataset).

As can be seen by comparing the mean average-precision (mAP) results in Table 1, the **MFSC** variants consistently improve upon their baselines, across the different 'hardness' sub-splits of the test-set. While previous work has focused on improving the quality of descriptors (whether handcrafted or learned) which are matched by a simple nearest-neighbor (NN) based assignment, we take a holistic approach to the matching problem, exploiting more information and specific properties of the task. For example, the meta-learning nature of MAML which allows for the descriptor (embedding) to adapt to the specific data at test time or the transductive nature of [47] and [16] which allows to treat the matching bijectivity.

4.2 Matching for Image Alignment

In this experiment we use the comprehensive FM-Bench [5] benchmark, following its evaluation protocols *verbatim*. It includes thousands of image matching pairs, over a diverse set of datasets, with the goal of evaluating components of geometry estimation pipelines for fundamental matrix estimation. The comparison is done by ablating each of the three main stages of a typical matching pipeline: (i) Local feature description and establishing initial correspondences across images ; (ii) Correspondence pruning ; and finally (iii) Fitting of a

Table 2: **Image Alignment on FM-Bench [5]**: CPC [40] dataset. **Bold** results are best 'recall', which refers to the percentage of image pairs whose matching error (Normalized SGD) was below 0.05; 'desc', 'fit' and 'prune' are short for descriptor, fitting method and pruning method; 'RSC' and 'LMS' are short for RANSAC and LMedS; See text for details on the additional measures: 'IR-m'/'IR' and 'corrs-m'/'corrs'.

desc	**fit**	**prune**	***recall***	***IR-m***	***IR***	***corrs-m***	***corrs***
SIFT	RSC	RT	28.2	48.1	67.2	415.3	60.5
		MFSL	**34.5**	53.4	82.0	352.0	54.0
	LMS	RT	45.5	48.1	75.4	415.3	208.3
		MFSL	**52.8**	53.4	81.8	352.0	177.0
HardNet++	RSC	RT	49.5	80.0	87.2	259.1	52.7
		MFSL	**52.4**	78.5	88.3	367.7	70.5
	LMS	RT	61.9	80.0	88.3	259.1	130.5
		MFSL	**64.3**	78.5	88.8	367.7	184.9

Table 3: **Person Re-Identification (Re-ID)** results on CUHK03 [21] and *Market-1501* [45]. Best results in **bold**. See text for details.

benchmark	CUHK03-det		CUHK03-lab		Market-1501	
network	***mAP***	***Rank-1***	***mAP***	***Rank-1***	***mAP***	***Rank-1***
MHN [7]	65.4	71.7	72.4	77.2	85.0	95.1
OSNet [46]	67.8	72.3	–	–	84.9	94.8
BDB [10]	73.6	76.4	76.7	79.4	86.2	94.5
MFSC-BDB	**75.8**	**77.3**	**80.4**	**89.8**	**87.0**	**95.2**

geometric model using a robust estimator.

We experiment with using the proposed MSFL matcher to replace the feature correspondence and pruning stages. To do so, we first extract features using a baseline feature extractor (SIFT [23] or HNet++ [25]). Note we did not fine tune HNet++ [25] here, since there is no train/test split. Next, we run a 1-shot-1-query-M-way FSC, where M is the number of features in image 1, the sets A and B are the interest points in images 1 and 2 respectively. We used the PTMAP [16] based ***Transd-FSC***, with mode='partial', to account for partial feature matching. Finally, a robust FM estimator is applied on the putative matches (RANSAC [13] or LMedS [24]). Lowe's ratio-test (RT), also known as "second nearest neighbor test", was used as the baseline for comparison, with the default threshold of 0.8.

Table 2 reports resutls on the CPC [40] dataset, which is the most challenging in [5]. We follow [5] and report: ***recall*** is the overall percentage (of all image pairs to match) of 'accurate' estimates - those for which the Normalized SGD error is less than a threshold of 0.05. ***IR-m*** and ***IR*** are the average (over matching pairs) of the inlier rates of obtained matches, before and after outlier rejection (that is part of the robust estimation stage). Finally, ***corrs-m*** and ***corrs*** are the number of correspondences obtained, before and after robust estimation. Note that ***recall*** is the main metric of [5], wherein MFSL shows a major boost. This can be attributed to several limitations of ratio-test [23] that is able to overcome: The ability to (i) avoid or resolve non-injective matches; (ii) allow a flexible (not predetermined) ratio between the first and second NNs; and (iii) pick matches that are not necessarily NNs.

4.3 Matching for Person Re-Identification (Re-ID)

In this section, we explore the possibility of applying MFSC to large-scale datasets by considering the person Re-Identification task [43]. Given a set of *query* images and a large set of *gallery* images, the task is to rank the similarities of each query against the entire gallery set. The ranking is typically done by learning specialized image features and comparing query features to gallery ones using Euclidean distances. MFSC is applied on pre-trained Batch DropBlock [10] resnet-50 features (termed BDB) and tested on the large-scale ReID benchmarks CUHK03 [21] (both 'detected' and 'labeled' versions) as well as the Market-1501 [45] set, reporting the conventional ***mAP*** (mean Average Precision) and ***Rank-1*** metrics. Note that we return the sorted ranking by probability

The results, where the ranking is obtained by returning the sorted probability indices rather than just the maximal one (line 2. in MFSC), are presented in Table 3. They show that the addition of MFSC clearly improves the baselines in both measures, across datasets. It can handle large-scale instances (with number of features in tens of thousands) and successfully improve performance measures in such retrieval-oriented tasks. Moreover, it successfully handles the more general setting (compared to the previous experiments) where the data is severely unbalanced (in number of images per individual) and matches are one-to-many, where we significantly improve the ***mAP*** measure that captures not only the best match.

Table 4: **MFSC components ablation**. **bold**/<u>underlined</u> results are best/second-best per column. See text for explanations.

method	feature	'easy'	'hard'	'tough'	*mean*
Nearest-Neighbor	Hnet++	72.2%	56.2%	37.9%	55.4%
	SOSNet	76.3%	68.4%	56.5%	67.1%
Sinkhorn	Hnet++	79.0%	66.5%	47.7%	64.4%
	SOSNet	81.0%	75.4%	65.2%	73.9%
MFSC	Hnet++	82.2%	71.6%	52.8%	68.8%
	SOSNet	<u>84.3%</u>	<u>80.0%</u>	<u>71.6%</u>	<u>78.6%</u>
MFSC (meta)	Hnet++	77.9%	73.0%	63.9%	71.6%
	SOSNet	**85.2%**	**81.1%**	**73.1%**	**79.8%**

4.4 Ablations

In Table 4 we present four different ablation levels of our method, over the HPatches matching task, for both the state-of-the-art HNet++ and SOSNet features: "Nearest-Neighbor" (NN) is the conventional closest-embedding matching; "Sinkhorn" is the result of applying Sinkhorn on the pairwise correlation matrix to produce a probability (assignment) matrix, showing a significant improvement over NN; "MFSC / (meta)" relate to our method, where the feature embeddings are pre-trained in the former or learnt in the latter. Our full approaches obtain major improvements across the board.

We also ran ablations on the main hyper-parameters of Algorithm ***Transd-FSC***, with the resulting *blue* plot mAP performance curves in Figure 4. These ablations were performed on the HPatches matching task, for different ranges of: k (main-loop count, step 4), α (center update weight, step 4 (c)) and λ (Sinkhorn entropy parameter). The vertical *red* lines depict the default PTMAP values that we worked with throughout the experiments. As can be seen, the original choices provide a good setting, with low sensitivity to the precise choice. We also used the default power-transform (PT) choice of $\beta = 0.5$, with $\beta = 1$ (i.e. no PT) when features are not non-negative. In addition, the orange curves in Figure 4 quantify the *injectivity* of the matching, by measuring the percentage of unique matches. Clearly, it is not fully correlated with accuracy, and can be tightly manipulated by the Sinkhorn entropy parameter. Interestingly, it is also affected by the other PTMAP parameters.

5 Conclusions

In this work, we propose a new perspective on set-to-set matching tasks, showing that they can be simply reduced to few-shot classification (FSC) tasks, for which a rich and well studied range of techniques and methodologies are available to be used. While the typical FSC methods were not meant to handle the (rather extreme) properties of the resulting tasks we obtain, we show that particular choices work well on a range of matching problems that arise in the applications of patch-matching, pose-estimation and person re-identification.

We have demonstrated, conceptually and empirically, that certain ideas in meta-learning and transductive inference are highly relevant to matching. Nevertheless, many aspects of this setting need to be further studied. For one thing, the solution we suggest is asymmetric (queries classified among classes), while the matching problem is symmetric by nature. In addition, matching between multiple sets could be possibly tackled in a more general way.

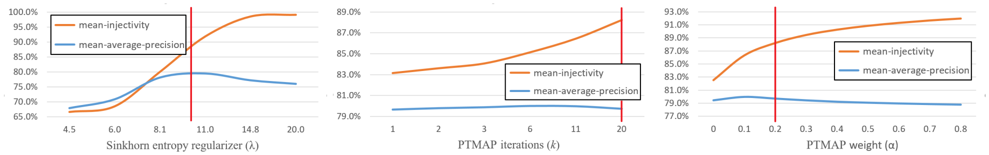

Figure 4: ***Transd-FSC*** **Hyper-parameter Ablations:** Sinkhorn entropy λ, PTMAP iterations k and weight α. Red line shows the choice we use, which is the default from [16].

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 2011.

[2] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. In *International Conference on Learning Representations (ICLR)*, 2019.

[3] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012.

[4] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, Tinne Tuytelaars, Jiri Matas, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (tPAMI)*, 2019.

[5] Jia-Wang Bian, Yu-Huan Wu, Ji Zhao, Yun Liu, Le Zhang, Ming-Ming Cheng, and Ian Reid. An evaluation of feature matchers for fundamental matrix estimation. In *British Machine Vision Conference (BMVC)*, 2019.

[6] JiaWang Bian, Wen-Yan Lin, Yasuyuki Matsushita, Sai-Kit Yeung, Tan-Dat Nguyen, and Ming-Ming Cheng. Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2019.

[8] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations (ICLR)*, 2018.

[9] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

[10] Zuozhuo Dai, Mingqiang Chen, Xiaodong Gu, Siyu Zhu, and Ping Tan. Batch dropblock network for person re-identification and beyond. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[11] Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *International Conference on Learning Representations (ICLR)*, 2020.

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

[13] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981.

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012.

[15] Kun He, Yan Lu, and Stan Sclaroff. Local descriptors optimized for average precision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] Yuqing Hu, Vincent Gripon, and Stéphane Pateux. Leveraging the feature distribution in transfer-based few-shot learning. In *International Conference on Artificial Neural Networks (ICANN)*, 2021.

[17] Gabriel Huang, Hugo Larochelle, and Simon Lacoste-Julien. Are few-shot learning benchmarks too simple? solving them without task supervision at test-time. *arXiv preprint arXiv:1902.08605*, 2019.

[18] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and Temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 2017.

[19] Simon Korman and Shai Avidan. Coherency sensitive hashing. *IEEE transactions on pattern analysis and machine intelligence (PAMI)*, 2015.

[20] Simon Korman, Mark Milam, and Stefano Soatto. Oatm: Occlusion aware template matching by consensus set maximization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[21] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[22] Wen-Yan Lin, Fan Wang, Ming-Ming Cheng, Sai-Kit Yeung, Philip HS Torr, Minh N Do, and Jiangbo Lu. Code: Coherence based decision boundaries for feature correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence (tPAMI)*, 2017.

[23] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision (ICCV)*, 1999.

[24] Desire L Massart, Leonard Kaufman, Peter J Rousseeuw, and Annick Leroy. Least median of squares: a robust method for outlier and model error detection in regression and calibration. *Analytica Chimica Acta*, 1986.

[25] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[26] Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Repeatability is not enough: Learning affine regions via discriminability. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[27] Rahul Mitra, Nehal Doiphode, Utkarsh Gautam, Sanath Narayan, Shuaib Ahmed, Sharat Chandran, and Arjun Jain. A large dataset for improving patch matching. *arXiv preprint arXiv:1801.01466*, 2018.

[28] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[29] Mattis Paulin, Matthijs Douze, Zaid Harchaoui, Julien Mairal, Florent Perronin, and Cordelia Schmid. Local convolutional features with unsupervised training for image retrieval. In *Proceedings of the IEEE international conference on Computer Vision (ICCV)*, 2015.

[30] Rodolfo Quispe and Helio Pedrini. Top-db-net: Top dropblock for activation enhancement in person re-identification. *25th International Conference on Pattern Recognition (ICPR)*, 2020.

[31] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations (ICLR)*, 2017.

[32] Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*. John wiley & sons, 2005.

[33] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[34] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[35] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of RGB-D SLAM systems. 2012.

[36] Yurun Tian, Bin Fan, and Fuchao Wu. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[37] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. Sosnet: Second order similarity regularization for local descriptor learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[38] Philip HS Torr and Andrew Zisserman. Mlesac: A new robust estimator with application to estimating image geometry. *Computer vision and image understanding (CVIU)*, 2000.

[39] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[40] Kyle Wilson and Noah Snavely. Robust global translations with 1DSFM. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014.

[41] Simon Winder, Gang Hua, and Matthew Brown. Picking the best daisy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[42] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *arXiv preprint arXiv:2001.04193*, 2020.

[43] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence (tPAMI)*, 2021.

[44] Sung Whan Yoon, Jun Seo, and Jaekyun Moon. Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. In *International Conference on Machine Learning (ICML)*, 2017.

[45] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[46] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[47] Imtiaz Masud Ziko, Jose Dolz, Eric Granger, and Ismail Ben Ayed. Laplacian regularized few-shot learning. In *International Conference on Machine Learning (ICML)*, 2020.