

Learning Dynamic Alignment via Meta-filter for Few-shot Learning

Chengming Xu^{*1} Yanwei Fu^{*1} Chen Liu¹
Chengjie Wang² Jilin Li² Feiyue Huang² Li Zhang¹ Xiangyang Xue¹
¹Fudan University ²YouTu Lab, Tencent

Abstract

Few-shot learning (FSL), which aims to recognise new classes by adapting the learned knowledge with extremely limited few-shot (support) examples, remains an important open problem in computer vision. Most of the existing methods for feature alignment in few-shot learning only consider image-level or spatial-level alignment while omitting the channel disparity. Our insight is that these methods would lead to poor adaptation with redundant matching, and leveraging channel-wise adjustment is the key to well adapting the learned knowledge to new classes. Therefore, in this paper, we propose to learn a dynamic alignment, which can effectively highlight both query regions and channels according to different local support information. Specifically, this is achieved by first dynamically sampling the neighbourhood of the feature position conditioned on the input few shot, based on which we further predict a both position-dependent and channel-dependent Dynamic Meta-filter. The filter is used to align the query feature with position-specific and channel-specific knowledge. Moreover, we adopt Neural Ordinary Differential Equation (ODE) to enable a more accurate control of the alignment. In such a sense our model is able to better capture fine-grained semantic context of the few-shot example and thus facilitates dynamical knowledge adaptation for few-shot learning. The resulting framework establishes the new state-of-the-arts on major few-shot visual recognition benchmarks, including miniImageNet and tieredImageNet.

1. Introduction

Deep learning models have excelled in many computer vision tasks such as image recognition [15, 39, 19] and ob-

^{*} indicates equal contributions. Li Zhang is the corresponding author.

Chengming Xu, Yanwei Fu, Chen Liu, Li Zhang and Xiangyang Xue are with School of Data Science, MOE Frontiers Center for Brain Science, and Shanghai Key Lab of Intelligent Information Processing, Fudan University. Email: {cmxu18, yanweifu, chenliu18, lizhangfd, yxxue}@fudan.edu.cn

Chengjie Wang, Jilin Li and Feiyue Huang are with YouTu Lab, Tencent. Email: {jasoncjwang, jeronlinli, garyhuang}@tencent.com

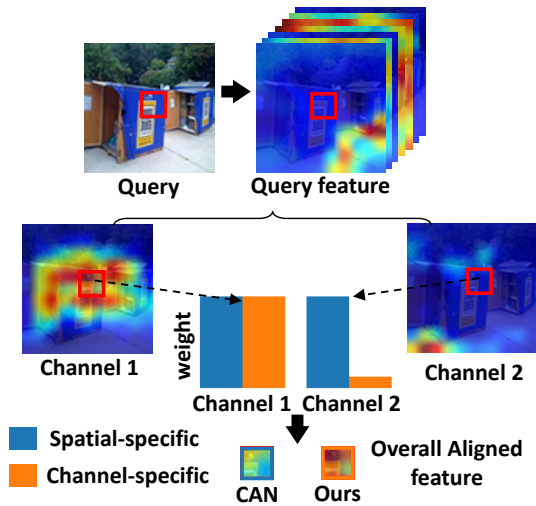


Figure 1. While position-specific alignment cannot eliminate the negative influence of badly-learned channels, channel-specific one can. CAN denotes Cross Attention Network [16] which is one of the position-specific alignment models in FSL.

ject detection [14, 46]. However, they highly rely on an avalanche of labeled training data and have difficulty transferring the learned knowledge to unseen categories. For example, an object detector trained on 80 categories of MSCOCO [23] would fail to detect a new class of *mouse*. This severely limits their scalability to open-ended learning of long tail categories in real-world. In contrast, learning from extremely constrained (*e.g.*, one or few) examples is an important ability for humans. For example, children have no problem forming the concept of “giraffe” by only taking a glance from a picture, or hearing its description as looking like a deer with a long neck.

Motivated by above observations, there has been a recent resurgence of research interest in few-shot learning (FSL) [18, 11, 43, 40, 42]. It aims to recognise new classes by adapting the learned knowledge with extremely limited few-shot (support) examples. The most naive baseline for few-shot classification is to learn a discriminative feature representation by deep convolution neural network. Query features are then assigned with the class label of the nearest support feature. As an alternative to that, learnable metric

like RelationNet [42] is proposed, where a binary classifier consisting of multiple layers of neural networks is utilised to calculate the similarity between two images. In such a framework, there is no exploration of more valuable information for specific query image when being classified with different support images, which leads to bad generalization ability. Although there are some recent works targeting feature alignment for FSL, such as CAN [16] and FEAT [48], which aggregates support knowledge to formalize an alignment function for query samples, we claim that these methods mainly have the following drawbacks. (1) *Roughness*. Due to limited knowledge in FSL, every channel would contribute to the prediction when comparing support feature to the query feature. Moreover, since the extracted features only have low spatial resolution, the information discrepancy among spatial positions is far smaller than that among channels for each query features. However, methods like CAN and FEAT do not focus on channel-level information. For example as shown in Fig. 1, while the spatial-specific alignment can set a large weight to the red box inside the target object, such a large weight is assigned to both well-learned and badly-learned channels. This would result in low signal in this region in the overall feature after alignment. In contrast, channel-specific alignment depends on the quality of each channel, thus being able to set low weight to channel 2, resulting in a better overall feature. (2) *Redundant matching*. There exists lots of redundancy in the support knowledge. For example, when classifying one position containing the target object, if there are several regions that also have this object, then comparing to one of them is enough for classification. Nonetheless, the existing methods utilize the whole support feature when aligning each query position, which is inefficient. (3) *Inflexible alignment*. The alignment strategy in these works only runs one time for all tasks. Thus, for those difficult ones, the alignment may be insufficient to appropriately embed the support knowledge into query feature.

Therefore to solve these problems, in this paper we propose a novel dynamic feature alignment strategy. In detail, we turn to dynamic filters [17, 52] for tackling this problem. We first predict a dynamic meta-filter with both position-specific and channel-specific filter weights based on small neighbourhood of each position. This filter can contain adequate information to inform the model of the most important regions and channels that need to be highlighted. Hence, applying this filter to align the query features will culminate in more effective representations for recognition. Meanwhile, using the neighborhood rather than the whole support feature directly decreases the number of knowledge source and the redundancy. To alleviate the problem resulted from fixed neighbor, we adopt a dynamic sampling strategy that all of the feature positions can be selected via learning an offset conditioned on the input few shot pairs.

Intuitively, this learned sampling allows the network to better capture position based semantic context of the few-shot example. To further make the alignment more adaptive to harder tasks, a direct way is to recursively apply the support knowledge to query feature through repetitive alignment. However, it is hard to control the extent of alignment by fixed hyper-parameter for various tasks. Consequently we modify the direct intuition of recursive alignment into an adaptive manner by using Neural Ordinary Differential Equation (ODE) which makes continuous the residual alignment procedure and takes an adaptive step size to get the final solution for the corresponding ODE. After achieving the aligned query feature, we utilize a meta-classifier to get the final prediction where the support knowledge is aggregated by unlearnable operations to form a classifier, which can avoid the adaptation problem that learnable classifiers are fully dependent on meta-train set and cannot adjust well for data in novel categories.

The contributions of this work are as follows:

1. We propose to learn a novel dynamic meta-filter for more effective and efficient feature alignment in FSL.
2. We introduce dynamic sampling and grouping strategy to further improve the flexibility and efficiency of basic dynamic meta-filter.
3. Neural ODE is, for the first time, leveraged to help model get better representation for FSL.

2. Related work

Few-shot recognition. Few-shot learning (FSL) is a surging research topic which aims to learn patterns with a set of data (base classes) and adapt to a disjoint set (new classes) with limited training data. Few-shot image classification is the one with most focus and researches. There are two main ways to tackle this problem. One is optimization-based methods [34, 9, 30, 21, 41], which firstly train a network with base class data, then finetune the classifier or the whole network with support data from unseen classes.

The metric-based method, on the other hand, is designed to solve FSL by applying an existing or learned metric on the extracted features of images. MatchingNet [43] adopts memory module to merge the information in each task and cosine distance as the metric to classify unseen data. ProtoNet [40] proposes the prototype as a simple representation of each category and adopts euclidean distance as the metric. RelationNet [42] uses network to learn the relation, which is taken as similarity, between input few-shot pairs. CAN [16], based on RelationNet, learns an attention module to highlight the correct region of interest to help better classification. The key dissimilarity between our model and the existing metric-based methods is that we propose to dynamically sample local context and based on that

to learn position and channel-dependent relationship which has never been explored before.

Dynamic Sampling. Sampling local context in a proper manner has been studied in the literature for a long time. Prior to the rise of deep learning, SIFT [27] and DPM [8] attempt to construct a good adaptive kernel. In the era of deep learning, convolutional kernel samples the neighbourhood of corresponding position in a uniform manner. For instance, a uniform 9-neighbourhood is sampled for a 3×3 kernel. Though this strategy paves the way for the success of the convolutional neural networks (CNNs), it is still limited in capturing position based semantic context. Our method is related to deformable convolution [4], where, instead of uniform sampling, a specific region is learned for each position. A fundamental difference between our model and the existing ones is that they only learn the offset dependent on the input feature while the filter weights are fixed for all inputs. In contrast, our model learns the offset to sample the neighbourhood of the feature position. The position and channel dependent filter is then predicted based on the sampled neighbourhood. Intuitively, this learned sampling allows the network to better capture position based semantic context of the few-shot example.

Input dependent weights. The basic idea of input dependent weights is to control the model weight not by directly optimizing but another learnable module. [17] developed an idea of “dynamic convolution”, that is predicting a dynamic convolutional filter for each feature position. HyperNet [13] builds additional module to generate input-dependent weight for a RNN. [1] firstly imports this idea to FSL. [11] proposes to generate final classifier weights based on pretrained classifier weights and the input information during test phase. [22] implants several new filters to the top layer while freezes other part of the model during test phase. The newly added weights is finetuned using support data. [33] aims to bridge the activation and classifier weights via additional sub-network. Unlike the above works that generate input-dependent or category-dependent weights, we focus on the local information and discrepancy between different positions and channels of query images. Our model can generate both position and channel specific filter weights which are especially beneficial for us to model the context of each feature position for target class, thus leading to more effective few-shot classification.

Neural Ordinary Differential Equation. Neural ODE was proposed by Chen et. al. in [3], which treats forward pass of a residual network as a discrete form of ODE. Under such condition, neural networks can be modified into neural ODEs where a time variable is introduced to control the output. There is a line of works targeting more efficient and robust neural ODE [28, 36, 5], and another line is the application of neural ODE to other tasks. For example, ODE²VAE [49] utilizes neural ODE for modeling tra-

jectory of high-dimensional data, and Vid-ODE [32] uses it in video generation. In this paper, neural ODE is, for the first time, introduced to few-shot learning to help learn a dynamic alignment between support and query features.

3. Problem formulation

The assumption of data split in FSL is different from common supervised learning. Suppose that we have two sets of data, meta-train set $\mathcal{D}_s = \{(\mathbf{I}_i, y_i), y_i \in \mathcal{C}_s\}$ and meta-test set $\mathcal{D}_t = \{(\mathbf{I}_i, y_i), y_i \in \mathcal{C}_t\}$ standing for collections of base class data and novel class data. \mathcal{C}_s and \mathcal{C}_t represent base and novel category sets respectively ($\mathcal{C}_s \cap \mathcal{C}_t = \emptyset$). The goal of FSL is to train a model on \mathcal{D}_s which is well generalized to \mathcal{D}_t . According to the commonly-used setting, we can access to few (*e.g.*, one or five) labelled data from each category of \mathcal{C}_t . These ground truth provide the supervision we can use to transfer the knowledge learned from base classes to novel classes.

We follow former methods [40, 42] to adopt an N -way K -shot meta-learning strategy. Here N denotes number of categories in one episode and K stands for number of samples for each category in support set. In detail, we sample N categories from \mathcal{C}_s for training and \mathcal{C}_t for testing, K instances each for these selected categories to construct a support set $\mathcal{S} = \{(\mathbf{I}_i^{supp}, y_i^{supp})\}$. Similarly we sample Q pieces of data from the N categories as query set $\mathcal{Q} = \{(\mathbf{I}_i^q, y_i^q)\}$, and $\mathcal{S} \cap \mathcal{Q} = \emptyset$.

4. Dynamic Alignment Network

Our model is illustrated in Fig. 2 where four sub-modules are used: feature extractor f_{emb} , dynamic alignment module f_d , meta-classifier f_{mc} and a global-classifier f_{gc} . Given each pair of support and query image $(\mathbf{I}^{supp}, \mathbf{I}^q)$, first we use f_{emb} to extract feature maps $X^{supp} = f_{emb}(\mathbf{I}^{supp})$, $X^q = f_{emb}(\mathbf{I}^q)$. The size of each feature map is $c \times h \times w$, where c, h, w indicates the number of channel, height and width, individually. The feature for each category can be represented as the average of all support feature maps in this class. For simplicity we will still use X^{supp} to stand for feature maps of each category. Next the dynamic alignment module f_d , which consists of a dynamic meta-filter and adaptive alignment, is used to align query feature with both position and channel-specific support knowledge. Finally, a confidence score of similarity between the adjusted query feature and corresponding support feature is provided by the meta-classifier f_{mc} . Meanwhile, a global classifier f_{gc} is also applied to the query feature to get a prediction of real category. All these modules are trained with a few-shot classification loss and a global classification loss. In the following context we describe our dynamic meta-filter by first explaining the naïve alignment model in Sec. 4.1 and then proposing and improving our dy-

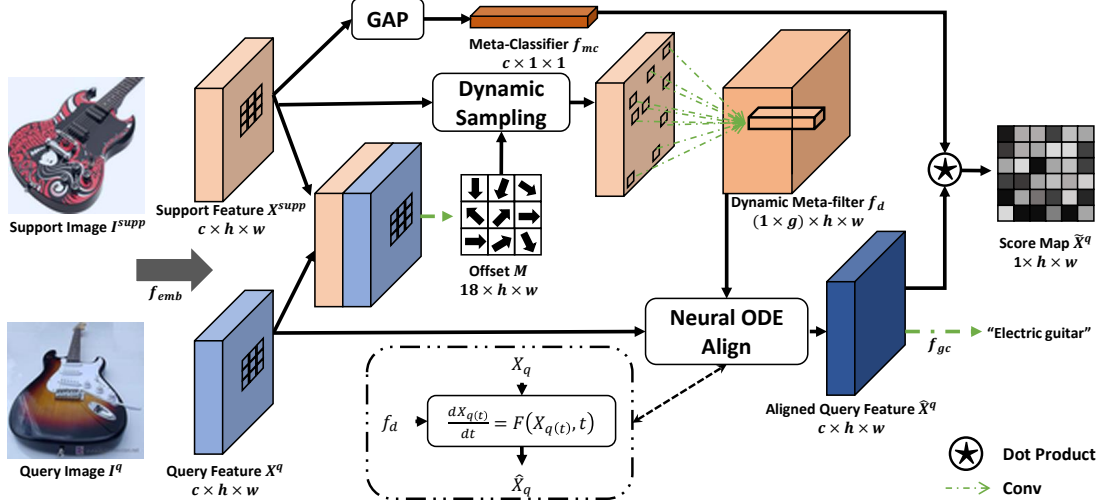


Figure 2. Schematic illustration of our proposed model for 1-shot task. Support and query features are extracted with a backbone network f_{emb} . Then two features are employed to predict an offset map using the local knowledge, which is used to dynamically sample the support feature to collect useful information for generating dynamic meta-filter for each query position. The filter is applied to align the query feature which is classified with a meta-classifier. GAP denotes global average pooling.

dynamic meta-filter from multiple points in Sec. 4.2. Then we complete the framework by introducing the meta-classifier (Sec. 4.3) and some implementation details (Sec. 4.4)

4.1. Naïve feature alignment

After extracting support and query features, several former studies show that applying an alignment or transformation to the query features based on support knowledge can help the model filter out feature part most beneficial to classification, thus improving the performance. A naive alignment can be written in the following form:

$$\hat{X}^q = g(X^q, A(X^q, \mathcal{X})) \quad (1)$$

where g is the aligning operation, A is alignment basis and \mathcal{X} is knowledge source for generating A . For CAN [16], g is multiplication, \mathcal{X} is single support feature and A is cross attention between two feature maps. For FEAT [48], g is summation, \mathcal{X} is whole support set and A is weighted average of \mathcal{X} with a self-attention module. The common ground of these two methods lies in two aspects: (1) Both methods omit channel-level alignment. CAN focuses on spatial-level attention while FEAT uses an instance-level one. (2) All available support knowledge is used. These two properties lead to the weaknesses including roughness, redundant matching and inflexible alignment as discussed in Sec. 1. To this end, we propose a new alignment framework in the form of dynamic meta-filter (DMF).

4.2. Dynamic Meta-filter for Adaptive Alignment

Concretely, given the support and query feature pair $\{X^{supp}, X^q\}$, instead of directly generating a filter from support features, we first apply a convolution layer f_ψ parameterized by ψ with kernel size 3 to the support feature

X^{supp} , which outputs a tensor $\psi(X^{supp}) \in \mathbb{R}^{(c \times k \times k) \times h \times w}$ where k denotes the kernel size of DMF. Then, for each position (i, j) we can have a tensor of size $c \times k \times k$:

$$f_d(i, j) = \sigma(\psi * \mathcal{B}_3(X_{:,i,j}^{supp})) \quad (2)$$

where \mathcal{B}_3 means a 3 width neighbor, $*$ denotes convolution, σ is the Sigmoid function used to control the scale. This $f_d(i, j)$ can be seen as c convolution filters with kernel size k . Convoluting $f_d(i, j)$ to the corresponding position of X^q with c groups leads to a refined query feature $\hat{X}_q \in \mathbb{R}^{c \times h \times w}$, where

$$\hat{X}_{:,i,j}^q = X_{:,i,j}^q + f_d(i, j) *_c \mathcal{B}_k(X_{:,i,j}^q) \quad (3)$$

where $*_c$ denotes convolution with group c . In this way, each channel is rescaled by a weight computed by DMF and the query feature is thus aligned according to the information collected from support feature in a both position-varying and channel-varying manner. Meanwhile, the source of support information for each query position is directly decreased from whole feature map to a small grid.

The dynamic meta-filter described above is both efficient and capable of dealing with fine-grained support knowledge, but drawbacks are also obvious: (1) The convolutional kernel used to generate DMF can only have a fixed kernel size, which would limit the receptive field, increasing the probability of failure to collect useful support information. (2) The DMF would be large and time consuming if the feature map has large dimension. (3) Despite a dynamic filter, the alignment is fixed, thus not flexible. Hence, we further propose three improvements to solve these problems.

Dynamic Sampling To help the model be more flexible to fetch any required features, we utilize a dynamic sampling strategy. In detail, the fixed convolution region in f_ψ

is replaced with a learnable one. Conditioned on the input support and query feature pair, a neighbourhood map $M \in \mathbb{R}^{9 \times h \times w}$ of each feature position is predicted with a convolution layer f_η of kernel size 5:

$$M = f_\eta(X^{supp} \| X^q) \quad (4)$$

where $\cdot \| \cdot$ denotes concatenation. Each row of the neighbourhood map $M_{:,i,j} \in \mathbb{R}^9$ denotes the regions containing useful information for generating the filter at (i, j) . This map is then used to dynamically sample knowledge from X^{supp} for each position to generate the DMF:

$$\begin{aligned} \tilde{\mathcal{B}}_3(X_{:,i,j}^{supp}) &= \text{Sample}(X_{:,i,j}^{supp}, M_{:,i,j}) \quad (5) \\ f_d(i, j) &= \sigma(\psi * \tilde{\mathcal{B}}_3(X_{:,i,j}^{supp})) \quad (6) \end{aligned}$$

Through dynamic sampling, we can increase the probability of exploring the useful positions while keeping a low volume of source knowledge. Among the many ways to implement the dynamic sampling, we imitate the deformable convolution (DeformConv) [4] where the neighborhood map is realized in the form of horizontal and vertical offsets $M \in \mathbb{R}^{18 \times h \times w}$ and sampling is based on the offset from each position. Note that [45] also adopts DeformConv in few-shot learning. However, they directly employ DeformConv to both the support and query features and map them into another latent space, which are then used to measure the correlation. No information interaction is conducted between support and query data. Compared to this method, the dynamic sampling in our model is predicted based on both support and query knowledge and used to generate a dynamic position and channel-dependent meta-filter which is then utilized to embed the knowledge from support images into query features.

Grouping Strategy To decrease the computation cost resulted from large number of filters, we make a simple assumption: similar information is shared among some groups of channels in feature maps. Attributed to that, we follow the existing works on group convolution [47] to gather these channels together to share one filter. Specifically, we equally segment the query feature map into g groups. The output channel of the meta-filter generator f_ψ is then modified from $c \times k \times k$ to $g \times k \times k$. Due to grouping the model only needs to generate a g -channel filter for each position, thus more lightweight. It is noteworthy that when $g = 1$, our DMF works in the same way as CAN and FEAT where all channels of each position are processed with the same weight. Such a strategy is shown to be less powerful than a larger g in our experiments, which proves our advantage against the other methods.

Adaptive Alignment The most direct solution for more difficult tasks is to recur the alignment process to get a more refined feature, such as in [44] where a learned spatial transformation is repeatedly used on the image features. By this

means, the query feature can be written as:

$$X_{t+1}^q = X_t^q + F(X_t^q, f_d), \quad t = 0, \dots, T-1 \quad (7)$$

$$\hat{X}^q = X_T^q \quad (8)$$

where F is the dynamic convolution. Such a method raises another problem — tuning the extra hyper-parameter of recursion depth T . In [44] the authors tried several choices of this hyper-parameter for one dataset. However, for different tasks, the depth variable should be varying depending on the level of complication. Therefore for FSL, it is inappropriate to use a fixed recursion depth on episodes containing various categories. As an alternative, we refer to the Neural ODE [3]. By gradually decreasing the time step t , the recursive residual equation 8 can be transformed from an Euler discretization form into an ODE:

$$\frac{dX^q(t)}{dt} = F(X^q(t), t) \quad (9)$$

Then \hat{X}^q is set as the solution of Eq. 9 given initial condition X^q . Such a form is better for two reasons. First, using modern ODE solvers such as Dormand-Prince method can get more accurate solution than Euler’s method. Second, by taking adaptive step size, the recursive depth is thus data-dependent, thus getting rid of hyper-parameter T .

4.3. Meta classifier

After we achieve a dynamically aligned query feature \hat{X}^q , we follow the CAN [16] to take a simple way to form the classifier which does not need any learnable parameters. Formally, we aggregate X^{supp} by global average pooling into $\bar{X}^{supp} \in \mathbb{R}^{c \times 1 \times 1}$, which contains the global information of \mathbf{I}^{supp} . We then use it as the weights for a convolutional filter, named meta-classifier (MC) f_{mc} , with c as the number of channel and 1×1 as kernel size. Applying this filter to query feature leads to a tensor $\tilde{X}^q \in \mathbb{R}^{1 \times h \times w}$. Through the calculation we directly compare the global support feature with local query feature on each corresponding channel. Therefore, if \mathbf{I}^q has the same category as \mathbf{I}^{supp} , \tilde{X}^q should have high value in most positions.

4.4. Implementation details

Objective function. Our loss function is defined in the same way as in CAN. For each input few-shot pair with all inputs and outputs $\{X^{supp}, y^{supp}, X^q, y^q, \hat{X}^q, \tilde{X}^q\}$, denote \mathcal{X} as the set of results from f_{mc} between \mathbf{I}^q and all support features, the objective function can be written as follow:

$$\mathcal{L} = \ell_f + \frac{1}{2} \ell_g \quad (10)$$

$$\ell_g = -(\log \text{softmax}(f_{gc}(\hat{X}^q)))^T y^q \quad (11)$$

$$\ell_f = \frac{1}{hw} \sum_{s,t} \log \frac{e^{-\tilde{X}_{s,t}^q}}{\sum_{\tilde{X} \in \mathcal{X}} e^{-\tilde{X}_{s,t}}} \cdot \mathbb{1}_{y^q = y^{supp}} \quad (12)$$

where ℓ_f is for the few-shot N -way classification, and ℓ_g is for the global many-shot $|\mathcal{C}_s|$ -way (e.g., 64 for *miniImageNet*) classification.

Network Structure. We use the same ResNet12 as in [48] for extracting image features. To enlarge the use of position-specific property of our DMF, we remove the last pooling layer of the backbone so that the spatial size of output feature map is doubled to 11×11 . Note that no extra parameters or capacity are introduced here. The convolution between our DMF and query features is implemented by first unfolding the query features along the spatial dimension and calculate the dot product between the unrolled features and DMF. The whole operation is highly efficient which will be shown in supplementary material.

5. Experiments

5.1. Datasets and setting

Datasets. Our experiments are conducted on two datasets. *miniImageNet* dataset [43], containing 600 images with each of the 100 categories, is a small subset of ImageNet. We follow the split in [34], where 64, 16, 20 classes are used for train, validation and test, respectively. *tieredImageNet* dataset [35] is a larger subset of ILSVRC-12 dataset. It consists of 34 categories with 779,165 images in total. These categories are further broken into 608 classes, where 351 classes are used for training, 97 for validation and 160 for testing. The size of images in *miniImageNet* is 84×84 and in *tieredImageNet*, 224×224 . Images in *tieredImageNet* are resized to 84×84 before training and testing.

Experimental setup. We empirically set the groups g as 64 for 1-shot tasks on both datasets, 160 for 5-shot *miniImageNet* and 320 for 5-shot *tieredImageNet*. Kernel size k is assigned as 1 for all tasks. Stochastic Gradient Descent (SGD) [2] with $5e-4$ weight decay is used to optimize our model. For *miniImageNet*, the initial learning rate is set as 0.35 and 0.05 for *tieredImageNet*. Cosine learning rate decay [26] is used along with SGD. Random cropping, horizontal flipping, color jittering and random erasing [53] are adopted for data augmentation during training, which is the same as in CAN [16]. We test 2000 episodes sampled from meta-test set for all experiments.

Evaluation benchmark. We report the accuracy and 95% confidence interval (CI) of 5-way 1-shot and 5-way 5-shot settings when comparing with the existing methods. For ablation study, only accuracy is reported.

Competitors. To show the efficacy of our model, we compare it with several previous methods for example Prototypical Network (ProtoNet) [40], RelationNet [42], MetaOptNet [20], Cross Attention Network (CAN) [16], etc. These models are chosen because they are among the best few-shot learning models and also the results with the same setting have been reported in the original paper.

5.2. Comparison with state-of-the-art

We compare our model with the competitors in Tab. 1, where the accuracies and 95% CI of 5-way 1-shot and 5-way 5-shot tasks on two datasets are shown. Note that different backbone structures are used among these models.

***miniImageNet* result.** As shown in the Tab. 1, on *miniImageNet*, all of the models with Conv4 are worse than ours, which in part results from the low capacity of the backbone. When comparing with models trained with the same backbone, the 1-shot accuracy of our model is still 0.98% higher than the best one, i.e., FEAT (66.78%). When comparing with models using WRN-28-10 which has larger capacity than ResNet12, our model is still outstanding. On 5-shot task, our model is 0.30% higher than DeepEMD which is the strongest competitor using ResNet12 backbone, and 1.54% higher than Robust dist++ which is the best 5-shot model with WRN backbone. It reflects that the ability of our model to deal with extremely limited data is better than most of the competitors, while ours is well-designed to draw the correct characteristics in common among limited images. Moreover, it is worth noting that the additional parameters of our model to the backbone are just two convolutional layers which is quite lightweight due to the group strategy. In contrast, most of these competitors, for example, TADAM and E³BM, have much more parameters. Also, some of the state-of-the-art methods for example FEAT and DeepEMD adopt pre-train to get a good initialization of the backbone for the meta-training stage so that their proposed modules used for few-shot recognition can be trained well. Compared with them, our proposed Dynamic Meta-filter does not depend on initialization, thus pretrain-free. In this way, our model can be trained with an ensemble of global classification loss and few-shot classification loss, which would be faster than those competitors.

***tieredImageNet* result.** Results in Tab. 1 show that on *tieredImageNet* we have 0.69% accuracy boost over the best competitor on 1-shot classification, which also proves the above conclusion. For 5-shot tasks, our model is 0.07% worse than DeepEMD. One possible reason is that *tieredImageNet* is a larger dataset with more base categories, hence the backbone itself can already learn a good representation with 5 support samples for each novel class. Moreover, DeepEMD receives marginally better results at the cost of massive training and inference time resulted from a more complicated classifier involving Quadratic Programming. In contrast, our framework is more succinct.

5.3. Model analysis

To further validate the effectiveness of our method, we conduct a series of ablation studies on *miniImageNet*. We first show some experimental proofs of the derivation of our method, then we show some other results on the hyperparameters. The results are shown in Tab. 2.

Model	Backbone	<i>miniImageNet</i>		<i>tieredImageNet</i>		
		1-shot	5-shot	1-shot	5-shot	
ProtoNet [40]	Conv4	49.42±0.78	68.20±0.72	53.31±0.89	72.69±0.74	
MatchingNet [43]		43.56±0.84	55.31±0.73	—	—	
RelationNet [42]		50.44±0.82	65.32±0.70	54.48±0.93	71.32±0.78	
MAML [9]		48.70±1.75	63.11±0.92	—	—	
Dynamic Few-shot [11]		56.20±0.86	72.81±0.62	—	—	
LEO [37]	WRN-28	61.76±0.08	77.59±0.12	66.33±0.05	81.44±0.09	
PPA [33]		59.60±0.41	73.74±0.19	—	—	
Robust dist++ [6]		63.28±0.62	81.17±0.43	—	—	
wDAE [12]		61.07±0.15	76.75±0.11	68.18±0.16	83.09±0.12	
CC+rot [10]		62.93±0.45	79.87±0.33	70.53±0.51	84.98±0.36	
FEAT [48]		65.10±0.20	81.11±0.14	70.41±0.23	84.38±0.16	
TapNet [50]	Res-12	61.65±0.15	76.36±0.10	—	—	
SNAIL [29]		55.71±0.99	68.88±0.92	—	—	
MetaOptNet [20]		62.64±0.61	78.63±0.46	65.99±0.72	81.56±0.53	
TADAM [31]		58.50±0.30	76.70±0.30	—	—	
DC [22]		62.53±0.19	78.95±0.13	—	—	
VFSL [7]		61.23±0.26	77.69±0.17	—	—	
CAN [16]		63.85±0.48	79.44±0.34	69.89±0.51	84.23±0.37	
FEAT [48]		66.78±0.20	82.05±0.14	70.80±0.23	84.79±0.16	
DeepEMD [51]		65.91±0.82	82.41±0.56	71.16±0.87	86.03±0.58	
E ³ BM [25]		63.80±0.40	80.10±0.30	71.20±0.40	85.30±0.30	
DSN-MR [38]		64.60±0.72	79.51±0.50	67.39±0.82	82.85±0.56	
Net-Cosine [24]		63.85±0.81	81.57±0.56	—	—	
Ours		Res-12	67.76±0.46	82.71±0.31	71.89±0.52	85.96±0.35

Table 1. 5-way few-shot accuracies with 95% confidence interval on *miniImageNet* and *tieredImageNet*.

DS	K=1	K=5	Pool	K=1	K=5
✓	67.76	82.71	✓	67.12	81.54
×	66.73	81.20	×	67.76	82.71

(a)

(b)

Group	K=1	K=5	Align	K=1	K=5
1	66.48	81.40	0	64.20	80.24
64	67.76	82.17	1	67.26	82.14
160	67.61	82.71	2	67.27	81.81
320	67.51	82.71	3	67.45	81.19
640	67.43	82.34	ODE	67.76	82.71

(c)

(d)

Table 2. Ablation Studies on *miniImageNet* 5-way tasks. We show 1-shot(K=1) and 5-shot(K=5) results. (a) **Dynamic Sampling**: Full model is compare with one without dynamic sample. (b) **Pooling**: we compare model trained with and without the last pooling layer in the Res-12 backbone. (c) **Groups**: we compare the number of groups in dynamic conv. (d) **Instantiation**: we try different structures based on our method, including models with no alignment, fixed number of alignment and ODE alignment.

Effectiveness of dynamic sampling We test models with and without dynamic sampling. As quantitative result in Tab. 2, the model with dynamic sampling is 1.03% better on 1-shot and 1.50% better on 5-shot than that without dy-

amic sampling. This shows that when directly using neighbor of each position to generate DMF, the model can only receive insufficient useful knowledge, leading to worse performance, which justifies the efficacy of this module. As an addition, we show the learned sampling strategy for some regions of support and query pairs in Fig. 3. This reflects that given specific query images, the dynamic sampling can successfully learn where to collect the knowledge, thus generating more useful filters for each region.



Figure 3. Visualisation of the learned dynamic sampling strategy. The red dots represent the sampled position in the support image that used to predict filter weights for the corresponding query feature position (green dot). With dynamic sampling, our model can utilize the truly informative positions to generate DMF.

Do we have to delete the pooling layer? The only difference between our backbone and the commonly-used ResNet12 is that we delete the last pooling layer, as discussed before. Tab. 2(b) shows the result when keeping this layer. The deletion of the pooling layer brings 0.64%

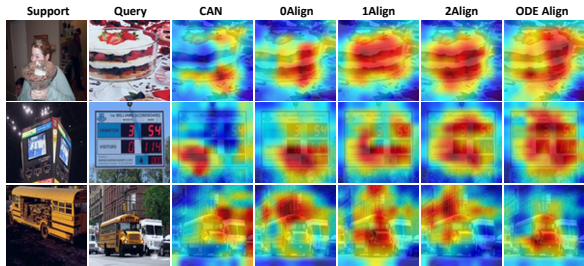


Figure 4. Comparison of aligned query feature maps from different methods for 5-way 1-shot tasks on *miniImageNet*.

improvement on 1-shot and 1.16% improvement on 5-shot. One reason is that when score map is larger, we can assign the loss function ℓ_f with more dense ground truth. Moreover, by abandoning this pooling layer, we can avoid information loss, thus helping learn better meta-filters.

Effectiveness of group convolution Our model is varied by different numbers of groups, as in Tab. 2(c). It reflects that as number of groups decreases, calculation between DMF and the corresponding region on query images gets simpler. When the number of groups equals to 1, this calculation degenerates from dot product between two vectors to element-wise product between scalar and vector, which is the same as in FEAT and CAN. This severely hurts the ability to express the whole information, leading to the worst performance. On the other hand, the profit brought by increasing number of groups saturates at different status, with 64 for 1-shot tasks and 160 for 5-shot tasks, and more groups would result in more computation consumption but along with performance drop. This means our assumption of shared information among channels in Sec. 4.2 holds. Since support knowledge only comes from one image for 1-shot tasks, the amount of information is less than that in 5-shot tasks. Hence, 64 filters each position are sufficient for handle 1-shot tasks, while the number of filters needs to be increased to 160 for 5-shot tasks to deal with more information.

Instantiation We compare four different kinds of instantiations of our propose methods as follows: (1) 0 align: The whole model is used without DMF, where the support and query features extracted with backbone network are directly delivered to meta-classifier. (2,3) 1/2 align: The DMF is used with fixed number of recursion. (4) ODE align: Neural ODE is used in place of recursive alignment, which is our final model. The results in Tab. 2(d) show that using one alignment can improve the model with no alignment by 1.28% on 1-shot and 0.77% on 5-shot, which verifies the motivation of using dynamic meta-filter for feature alignment. Furthermore, an interesting fact is that while using more alignments can slightly improve the performance on 1-shot tasks, the 5-shot accuracy dramatically decreases, with a 0.95% gap between models with 1 and 3 alignments. This phenomenon reflects the inflexibility of fixed number recursive alignment when dealing with different types of

tasks. Compared to this, using ODE align can boost the performance on both 1-shot and 5-shot tasks. Similar results are shown in Fig. 4, where we present three pairs of support and query images and the corresponding query feature maps from the above instantiations of our model along with CAN. We find that (1) With our dynamic alignment, the results are generally better than the ones generated by CAN and model with no alignment. Our DMF is able to focus on the target object, which suggests the efficacy of our method. (2) The model with one and two alignments can only handle different data. While both models can correctly align the image in the first row, the model trained with two alignments can have better representation than that with one alignment in the second row, and it is opposite in the third row. In contrast, the ODE align can correctly highlight the query feature in all circumstances, which proves our claim of the requirement of adaptive alignment.

5.4. Further Discussion

Our proposed dynamic meta-filter works well with the adaptive alignment in several few-shot learning datasets as shown. However, it cannot be ignored that some extreme conditions would weaken our method. For example, when the target object appears in very different parts of support and query images, e.g., upper left and lower right, it would be hard for our model to learn such an alignment. Even though such a case is too extreme to hurt the performance, finding a remedy for that will be a potential future work.

6. Conclusion

We propose to learn a novel class recognition network for few-shot learning with a novel dynamic meta-filter generated from the few-shot inputs. We dynamically sample a relevant neighbour for each feature position of few shot input and further predict position-specific and channel-specific filter weights based on the sampled neighbourhood to facilitate novel class recognition. This formulation is able to better capture position based semantic context of the few-shot example and thus enjoys better dynamical knowledge adaptation for few-shot learning. This is demonstrated by the fact that we establish new state-of-the-arts on major benchmarks *miniImageNet* and *tieredImageNet*.

Acknowledgement. This work was supported in part by NSFC Projects (U62076067), Science and Technology Commission of Shanghai Municipality Projects (19511120700, 19ZR1471800), Shanghai Research and Innovation Functional Program (17DZ2260900), Shanghai Municipal Science and Technology Major Project (2018SHZDZX01) and ZJLab.

References

- [1] Luca Bertinetto, João F Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. In *Adv. Neural Inform. Process. Syst.*, 2016. [3](#)
- [2] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. 2010. [6](#)
- [3] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, pages 6571–6583, 2018. [3](#), [5](#)
- [4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Int. Conf. Comput. Vis.*, 2017. [3](#), [5](#)
- [5] Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural odes. In *Advances in Neural Information Processing Systems*, pages 3140–3150, 2019. [3](#)
- [6] Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Diversity with cooperation: Ensemble methods for few-shot classification. In *Int. Conf. Comput. Vis.*, 2019. [7](#)
- [7] Nanyi Fei, Zhiwu Lu, Yizhao Gao, Jia Tian, Tao Xiang, and Ji-Rong Wen. Meta-learning across meta-tasks for few-shot learning. 2019. [7](#)
- [8] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009. [3](#)
- [9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. [2](#), [7](#)
- [10] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *Int. Conf. Comput. Vis.*, 2019. [7](#)
- [11] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. [1](#), [3](#), [7](#)
- [12] Spyros Gidaris and Nikos Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. [7](#)
- [13] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016. [3](#)
- [14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Int. Conf. Comput. Vis.*, 2017. [1](#)
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. [1](#)
- [16] Ruibing Hou, Hong Chang, MA Bingpeng, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *Adv. Neural Inform. Process. Syst.*, 2019. [1](#), [2](#), [4](#), [5](#), [6](#), [7](#)
- [17] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *Adv. Neural Inform. Process. Syst.*, 2016. [2](#), [3](#)
- [18] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. 2015. [1](#)
- [19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Adv. Neural Inform. Process. Syst.*, 2012. [1](#)
- [20] Kwonjoon Lee, Subhansu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. [6](#), [7](#)
- [21] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017. [2](#)
- [22] Yann Lifchitz, Yannis Avrithis, Sylvaine Picard, and Andrei Bursuc. Dense classification and implanting for few-shot learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. [3](#), [7](#)
- [23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, 2014. [1](#)
- [24] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. *arXiv preprint arXiv:2003.12060*, 2020. [7](#)
- [25] Yaoyao Liu, Bernt Schiele, and Qianru Sun. An ensemble of epoch-wise empirical bayes for few-shot learning. In *European Conference on Computer Vision*, pages 404–421. Springer, 2020. [7](#)
- [26] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. [6](#)
- [27] David G Lowe. Object recognition from local scale-invariant features. In *Int. Conf. Comput. Vis.*, 1999. [3](#)
- [28] Stefano Massaroli, Michael Poli, Jinkyoo Park, Atsushi Yamashita, and Hajime Asama. Dissecting neural odes. *arXiv preprint arXiv:2002.08071*, 2020. [3](#)
- [29] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*, 2017. [7](#)
- [30] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. [2](#)
- [31] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Adv. Neural Inform. Process. Syst.*, 2018. [7](#)
- [32] Sunghyun Park, Kangyeol Kim, Junsoo Lee, Jaegul Choo, Joonseok Lee, Sookyung Kim, and Edward Choi. Vid-ode: Continuous-time video generation with neural ordinary differential equation. *arXiv preprint arXiv:2010.08188*, 2020. [3](#)
- [33] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. [3](#), [7](#)
- [34] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *Int. Conf. Learn. Represent.*, 2017. [2](#), [6](#)

- [35] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. 2018. [6](#)
- [36] Yulia Rubanova, Ricky TQ Chen, and David Duvenaud. Latent odes for irregularly-sampled time series. *arXiv preprint arXiv:1907.03907*, 2019. [3](#)
- [37] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018. [7](#)
- [38] Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi. Adaptive subspaces for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4136–4145, 2020. [7](#)
- [39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. [1](#)
- [40] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Adv. Neural Inform. Process. Syst.*, 2017. [1](#), [2](#), [3](#), [6](#), [7](#)
- [41] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. [2](#)
- [42] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. [1](#), [2](#), [3](#), [6](#), [7](#)
- [43] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Adv. Neural Inform. Process. Syst.*, 2016. [1](#), [2](#), [6](#), [7](#)
- [44] Wanglong Wu, Meina Kan, Xin Liu, Yi Yang, Shiguang Shan, and Xilin Chen. Recursive spatial transformer (rest) for alignment-free face recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3772–3780, 2017. [5](#)
- [45] Ziyang Wu, Yuwei Li, Lihua Guo, and Kui Jia. Parn: Position-aware relation networks for few-shot learning. In *Int. Conf. Comput. Vis.*, 2019. [5](#)
- [46] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. *arXiv preprint arXiv:1909.13226*, 2019. [1](#)
- [47] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. [5](#)
- [48] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8808–8817, 2020. [2](#), [4](#), [6](#), [7](#)
- [49] Cagatay Yildiz, Markus Heinonen, and Harri Lahdesmaki. Ode2vae: Deep generative second order odes with bayesian neural networks. In *Advances in Neural Information Processing Systems*, pages 13412–13421, 2019. [3](#)
- [50] Sung Whan Yoon, Jun Seo, and Jaekyun Moon. Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. 2019. [7](#)
- [51] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover’s distance and structured classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12203–12213, 2020. [7](#)
- [52] Li Zhang, Dan Xu, Anurag Arnab, and Philip HS Torr. Dynamic graph message passing networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3726–3735, 2020. [2](#)
- [53] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017. [6](#)