Learning Cross-Image Object Semantic Relation in Transformer for Few-Shot Fine-Grained Image Classification

Bo Zhang†
Fudan University
Shanghai AI Laboratory, China
bo.zhangzx@gmail.com

Jiakang Yuan†
Fudan University
China
jkyuan18@fudan.edu.cn

Baopu Li
Oracle Health and AI
USA
baopu.li@oracle.com

Tao Chen‡
Fudan University
China
eetchen@fudan.edu.cn

Jiayuan Fan
Fudan University
China
jyfan@fudan.edu.cn

Botian Shi
Shanghai AI Laboratory
China
friskit.china@gmail.com

ABSTRACT
Few-shot fine-grained learning aims to classify a query image into one of a set of support categories with fine-grained differences. Although learning different objects’ local differences via Deep Neural Networks has achieved success, how to exploit the query-support cross-image object semantic relations in Transformer-based architecture remains under-explored in the few-shot fine-grained scenario. In this work, we propose a Transformer-based double-helix model, namely HelixFormer, to achieve the cross-image object semantic relation mining in a bidirectional and symmetrical manner. The HelixFormer consists of two steps: 1) Relation Mining Process (RMP) across different branches, and 2) Representation Enhancement Process (REP) within each individual branch. By the designed RMP, each branch can extract fine-grained object-level Cross-image Semantic Relation Maps (CSRM)s using information from the other branch, ensuring better cross-image interaction in semantically related local object regions. Further, with the aid of CSRMs, the developed REP can strengthen the extracted features for those discovered semantically-related local regions in each branch, boosting the model's ability to distinguish subtle feature differences of fine-grained objects. Extensive experiments conducted on five public fine-grained benchmarks demonstrate that HelixFormer can effectively enhance the cross-image object semantic relation matching for recognizing fine-grained objects, achieving much better performance over most state-of-the-art methods under 1-shot and 5-shot scenarios. Our code is available at: https://github.com/JiakangYuan/HelixFormer.

CCS CONCEPTS
• Computing methodologies → Visual content-based indexing and retrieval: Matching.

KEYWORDS
Cross-Image Object Semantic Relation, Transformer, Few-Shot Fine-Grained Image Classification

ACM Reference Format:
Bo Zhang, Jiakang Yuan, Baopu Li, Tao Chen, Jiayuan Fan, and Botian Shi. 2022. Learning Cross-Image Object Semantic Relation in Transformer for Few-Shot Fine-Grained Image Classification. In Proceedings of the 30th ACM International Conference on Multimedia (MM ’22), Oct. 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3547961

1 INTRODUCTION
The success of Deep Neural Networks (DNNs) [15, 17, 56], [32, 67] largely owes to the completeness of training data, which means that the collected data should be carefully annotated, low level noised, and sufficiently large. But in many real scenarios that need professional labeling knowledge, such as fine-grained image classification for different kinds of birds, it is generally hard to access a large number of label-rich training samples.

One solution to alleviate the over-dependence of DNNs on label-rich training data is Few-Shot Learning (FSL) [2, 4, 31, 39, 42, 61, 62], which aims to mine the transferable meta-knowledge from the base classes so that DNNs can utilize such knowledge to easily recognize new classes given only a few training examples. However, the main challenge for the FSL [2, 4, 31, 39, 42, 61, 62] is that learning to classify from few examples having limited representative capability inevitably brings the overfitting issue. Therefore, researchers mainly focus on leveraging meta-learning technology [13, 39, 41, 43, 48, 59] to deal with the FSL problem. However, the above-mentioned FSL methods focus to classify coarse-grained generic object categories, which are less suitable to address the few-shot fine-grained classification task [29, 57, 58, 71], that requires to emphasize the local feature variations or subtle feature differences.

Inspired by the meta-learning success for generic object classification, some researchers [9, 12, 18, 19, 27–29, 44–46, 50, 57, 58, 71] start to extend the study of FSL from generic object classes to fine-grained classes, where the main challenge is that fine-grained recognition is more dependent on mining spatially local discriminative parts of an input image, rather than global features extracted by generic meta-learning models such as prototypical networks [41]. As illustrated in Fig. 1(a), many few-shot fine-grained
works mine discriminative parts of the whole image based on the attention mechanism [71], feature map reconstruction [57], and feature-level spatial alignment [58]. However, these methods fail to leverage cross-image object semantic relations between the training examples (denoting the support images) and test examples (denoting the query images).

As a matter of fact, to recognize a novel class’s samples, humans tend to compare the local semantic parts’ differences between the ever-seen and newly-seen objects, so that the subtle feature differences can be identified using only a few examples. Thus, benefiting from such a fact, we are encouraged to model the cross-image object semantic relation to find the discriminative local regions in the Transformer, as illustrated in Fig. 1(b). We further give an in-depth analysis on how to design such a cross-attention mechanism under both few-shot and fine-grained scenarios via the following two aspects.

Firstly, under the FSL setting, we consider that the cross-image object semantic relation calculated from the support to query image, should be consistent with that calculated from the query to support image, namely, cross-image object semantic relation consistency, which motivates us to use a symmetrical structure to model the cross-image object semantic relations. Moreover, similar to humans that need to distinguish from multiple relations between a target object and its various neighbors to determine its belonged category, cross-image object semantic relation should be modeled as a parts-to-parts (many-to-many) matching process, that involves several attention heads of a Transformer running in parallel with each head focusing on a certain part.

Secondly, such cross-image semantic relation learning should not only consider the few-shot scenario, but also consider the fine-grained setting which requires to strengthen subtle local features’ discriminability. Different fine-grained objects in real world may present arbitrary poses, scales and appearances, causing different similarity in both global appearance and various local regions, hurting the classification performance. This challenge may be alleviated if the learned cross-image semantic relation can serve to emphasize those discriminative local features by integrating into a well-performed fine-grained baseline model, and enhance the representation learning process for those discovered semantically-similar regions.

In view of the above, we propose a Transformer-based double-helix model, namely HelixFormer to solve the few-shot fine-grained image classification task. Benefiting from the self-attention mechanism of Transformer, the HelixFormer exploits the multi-head key-query-value outputs to do interaction between different images, and predicts their object-level semantic relation. HelixFormer is mainly composed of a cross-image object Relation Mining Process (RMP) across the support-query branches, and a feature Representation Enhancement Process (REP) within each single branch. Specifically, we feed the input image pairs into the support and query branches respectively. First, by treating the input feature patches (extracted by the Convolution Neural Networks (CNNs) based backbone such as Conv-4 [43]) as tokens and considering the cross-branch token information interaction, RMP produces Cross-image Semantic Relation Maps (CSRMs) for each branch in a bidirectional and symmetrical manner, which ensures the cross-image relation consistency. Meanwhile, we formulate the multi-head cross-attention mechanism as modeling many-to-many inter-object semantic relations, where one object would interact with multiple other objects. Second, the above CSRMs encode image patch-level semantic relation between different fine-grained objects, and further help the subsequent REP to enhance the learning of the feature representations within either the support or query feature encoding branch, which boosts the baseline model’s ability to distinguish subtle feature differences of fine-grained objects.

The main contributions of this work can be summarized as follows:

1) We propose a novel HelixFormer architecture, leveraging on the cross-image object semantic relation learning at patch level, to perform the few-shot fine-grained learning. To our knowledge, this is the first work to introduce semantic relation mining in the Transformer model for few-shot fine-grained task.

2) To ensure the semantic relation consistency between a pair of support-query images in FSL, we design a double-helix RMP structure that can generate consistent patch-level CSRMs in different branches. Furthermore, with the aid of CSRMs, we develop a REP to enhance the feature learning for those semantically-similar regions presenting subtle feature differences.

3) We have conducted extensive experiments on five few-shot fine-grained benchmarks. Both qualitative and quantitative results validate that the HelixFormer can effectively learn the cross-image object semantic relations, and further utilize such relations to enhance the model’s generalization ability.

2 RELATED WORKS

Few-Shot Fine-Grained Learning (FSFGL). Recent FSL works [1, 23, 25, 37, 63, 65, 68] can be roughly categorized into three types: 1) Optimization-based methods [11, 38] that focus on learning good initialization parameters in order to quickly adapt the few-shot model to novel classes; 2) Metric-based methods [35, 41, 43, 48]
that aim to design a distance metric, so that the few-shot model can learn the semantic relation between different input images; 3) Data augmentation-based methods [6, 14, 26, 36, 54, 70] that produce new samples to enlarge the training set for model training. Recently, inspired by the rapid development of meta-learning, researchers [9, 12, 19, 27–29, 46, 50, 57, 58, 71] start to explore the generalization ability of FSL model on novel fine-grained sub-classes where only a few training examples are given. For example, a multi-attention meta-learning method [71] is proposed to learn diverse and informative parts of fine-grained images. Besides, the work [57] tackles the FSFGL problem from the perspective of reconstructing the query image to learn a classifier. More recently, the work [58] tries to increase the fine-grained classification accuracy via long-shot-range spatial alignment between support and query features. Motivated by these works in the FSFGL community, we further extend the study of FSFGL to a Transformer-based structure, and investigate its effectiveness in strengthening the support-query relation matching process only given a few samples.

Cross-attention Models. In this part, we review existing cross-attention works [3, 20, 21, 30, 52, 55, 60] and find that they are mainly based on the attention modeling of cross-scale features [3, 30, 52], cross-modality relationships [55, 66], joint spatio-temporal information [21], and inner-image multi-patches [20] to capture the intra-object relations and contextual information. Different from the above methods, our work aims to exploit the cross-image object semantic relations (i.e. finding the discriminative local spatial regions between objects that are helpful for fine-grained recognition) to address the FSFGL issue. On the other hand, there are only two works [16, 72] employing cross-attention mechanism to perform the FSL task. Our work differs from the two works as follows: 1) We study a symmetrical Transformer-based structure, which fully considers the symmetry property between support and query images in FSL, and imposes a cross-image object semantic relation consistency between support-query and query-support matching process; 2) We develop a two-step relation matching process (a two-branch relation mining process and a representation enhancement process), which has the merit of improving the baseline model’s ability to distinguish the subtle local feature differences.

Transformer for Vision Tasks. Vision Transformer (ViT) [10] shows promising performance on a variety of tasks including image classification [10], object detection [7, 51], segmentation [7, 51], and pose estimation [66]. The goal of ViT is to model long-range dependencies across different input sequence elements (or input tokens), by splitting an input image into a sequence of image patches with size of 16×16 pixels. More recently, to reduce the computational cost, many researchers incorporate the multi-scale branch into Transformer, via inserting depth-wise convolutions into the self-attention [66] or exploiting the multi-resolution parallel structure [33]. The above works attempt to model the global self-attention within an input image through the single-branch structure, while our HelixFormer tries to identify the local patch-level cross-attention between different input images. Besides, CrossTransformer [8] and CrossViT [3] are recently developed Transformer-based dual-branch structures, where the CrossTransformer utilizes the dual-branch structure to achieve a coarse-grained spatial correspondence, and CrossViT feeds the image patches of different sizes into two separate branches to extract multi-scale information. We would like to emphasize that, compared with the above dual-branch network structure, our HelixFormer is actually a double-helix dual-branch structure, which means that the cross-object semantic relations from the support to query branch and vice versa are symmetric and complementary, ensuring the semantic consistency assumption of relation pairs.

3 THE PROPOSED METHOD

The overall framework of the proposed HelixFormer is illustrated in Fig. 2. For easy understanding, we first give the problem formulation and the episodic training strategy for Few-Shot Learning (FSL) task. Then we introduce the proposed HelixFormer and discuss its superiority over several variants of existing cross-attention models. Finally, we give the overall objectives and cross-attention learning strategy of our model.

3.1 Preliminaries

FSL Setting. Given a set of base classes $D_{\text{base}}$ and a CNN-based feature embedding network (or backbone) $F$, the purpose of few-shot learning is to learn a task-agnostic $F$ on $D_{\text{base}}$ via an episodic task sampling strategy, so that the $F$ can be generalized to novel classes $D_{\text{novel}}$, where $D_{\text{base}} \cap D_{\text{novel}} = \emptyset$. For a typical FSL setting, each episodic task represents an $N$-way $K$-shot classification task, where both support set $S$ and query set $Q$ are sampled from the same $N$ classes. During the meta-training stage, each episodic task is sampled from the base classes $D_{\text{base}}$.

Two-branch Baseline Introduction. Inspired by the two-branch network structure to learn semantic representations in the relation network (RN) [43], we employ the RN as our baseline model. As illustrated in Fig. 2, given an input image pair $(x_5, x_Q)$, the RN first produces the high-level semantic feature pairs $(f_5, f_Q)$ via a convolution-based backbone $F$, and then, a classification network $H$ is used to predict whether the query image $x_Q$ has the same class label with the $n$-th class support image $x_{n,S}$. Thus, the loss function of our baseline model $L_{\text{cls}}(F, H; x)$ can be formulated as follows:

$$L_{\text{cls}}(F, H; x) = \sum_{(x_Q, y_Q) \in Q} \log P(y_Q = n|x_Q; x_S)$$

$$= \sum_{n \in N} \exp(H(F(x_{n,S}), F(x_Q)))$$

(1)

where $P(x_{n,S}, F(x_Q))$ denotes the concatenation between $F(x_{n,S})$ and $F(x_Q)$, and $y_Q$ is the label of query image. During the meta-test or model evaluation stage, each episodic task is sampled from the novel classes $D_{\text{novel}}$, given the prototype of $K$ labeled support images of the $n$-class $x_{n,S}$. Note that the label $y_Q$ is available for model training on $D_{\text{base}}$, but it can only be used for model evaluation on $D_{\text{novel}}$.

3.2 HelixFormer

The purpose of this work is to capture the cross-image object semantic relations for improving the generalization ability of fine-grained model. By means of the multi-head attention module in Transformer, a many-to-many matching process of semantically related regions can be established. Note that just a single HelixFormer is sufficient in capturing the cross-attention, and please refer to our supplementary material for the study of stacking HelixFormer.

Learning Cross-Image Object Semantic Relation in Transformer for Few-Shot Fine-Grained Image Classification MM ’22, October 10–14, 2022, Lisboa, Portugal
Figure 2: The overview of the proposed HelixFormer, which takes a pair of support-query features \((f_S, f_Q)\) extracted by a CNN backbone as its input, and outputs a pair of \((\hat{f}_S, \hat{f}_Q)\) for doing the subsequent object classification. Note that for simplicity, we omit the multi-head attention in this figure.

**Bidirectional Relation Mining Process.** Given a pair of images \((x_S, x_Q)\) sampled from the \(S\) and \(Q\) respectively, the backbone \(F\) is first used to extract a pair of high-level features \((f_S, f_Q)\) where \(f_S = F(x_S) \in \mathbb{R}^{C \times H \times W}\), and \(C\) denotes the channel number, \(H\) and \(W\) are the height and width of the features, respectively. Although the feature pairs \((f_S, f_Q)\) contain rich semantics, they lack interaction from each other, and do not consider the cross-object semantic relations between support-query images.

To fully encode the relations between the two branches, we treat the feature maps \(f \in \mathbb{R}^{C \times H \times W}\) as a sequence of \(H \times W\) tokens, with each token having \(C\) channels, which can be formulated as \(f = [f^1, f^2, ..., f^{H \times W}]\), where \(f^i \in \mathbb{R}^C\).

In detail, given the token embedding with weight parameters \(W^e_S, W^k_S, W^v_S\) for support branch \(S\), and parameters \(W^e_Q, W^k_Q, W^v_Q\) for query branch \(Q\), the query vector \(e^i\), key vector \(k^i\), and value vector \(v^i\) can be calculated as follows:

\[
\begin{align*}
    e^i &= W^e_S f_S^i \\
    k^i &= W^k_S f_S^i \\
    v^i &= W^v_S f_S^i \\
    e^i_Q &= W^e_Q f_Q^i \\
    k^i_Q &= W^k_Q f_Q^i \\
    v^i_Q &= W^v_Q f_Q^i
\end{align*}
\]

(2)

where for avoiding the confusion of symbol definition, \(e^i\) denotes the query vector in the RMP, and \(Q\) represents the query branch for the whole pipeline. Besides, according to the ablation studies in Sec. 4.4, HelixFormer employs the convolution-based token embedding and removes the position embedding, since local spatial information has been encoded in the convolution feature maps. Further, we achieve the RMP by a symmetrical cross-attention from two directions: 1) We obtain support branch-related features using the value vector \(v^i_Q\) from another branch \(Q\), formulated as \(Q \rightarrow S\); 2) We also obtain query branch-related features using the value vector \(v^i_S\) of the support branch \(S\), formulated as \(S \rightarrow Q\).

For \(Q \rightarrow S\) direction, let \(A_{Q,S} \in \mathbb{R}^{H \times W \times H \times W}\) denote the matrix of attention scores obtained via the matrix multiplication as follows:

\[
A_{Q,S} = K_Q E_Q^T
\]

(3)

where \(K_Q = [k_{Q,1} \ldots k_{Q,H \times W}] \in \mathbb{R}^{H \times W \times C}\) and \(E_S = [e_{S,1} \ldots e_{S,H \times W}] \in \mathbb{R}^{H \times W \times C}\), which can be obtained using Eq. 2. Note that the designed token embedding way does not change the channel number for each input token \(f^i \in \mathbb{R}^C\). Moreover, to perform normalization for attention scores in Transformer and find the semantically related regions according to clues from another branch \(Q\), a softmax layer with a scaling factor is employed as follows:

\[
R_{Q,S} = \text{Softmax}(A_{Q,S}/\sqrt{C}) V_Q
\]

(4)

where \(V_Q = [v_{Q,1} \ldots v_{Q,H \times W}] \in \mathbb{R}^{H \times W \times C}\), and \(R_{Q,S} \in \mathbb{R}^{H \times W \times C}\) represents the Cross-image Semantic Relation Maps (CSRMs), encoding the patch-level semantic relations from query to support branch. Then, the CSRMs are reshaped to the same dimension as the backbone features \(f_S \in \mathbb{R}^{C \times H \times W}\), in order to enhance the semantically similar backbone features in the REP operation.

For \(S \rightarrow Q\) direction, the CSRMs \(R_{S,Q} \in \mathbb{R}^{H \times W \times C}\) also can be easily obtained by performing a symmetrical process described in Eqs. 3 and 4, which can be written as follows:

\[
A_{S,Q} = K_S E_Q^T
\]

\[
R_{S,Q} = \text{Softmax}(A_{S,Q}/\sqrt{C}) V_S
\]

(5)

**Representation Enhancement Process.** Based on the above RMP, bidirectional semantic relations of support-to-query features and query-to-support features have been symmetrically encoded into the matrix of attention scores, from both directions \(A_{Q,S}\) and \(A_{S,Q}\), so that these cross-image object semantically-similar parts can be first found. In this part, we design a REP that can further guide the classification network to learn how to distinguish these semantically similar features obtained by the RMP.

Given the high-level features \((f_S, f_Q)\) learned from the CNNs-based backbone, and the CSRMs \((R_{Q,S}, R_{S,Q})\) calculated from the \(Q \rightarrow S\) and \(S \rightarrow Q\), the REP can be formulated as follows:

\[
\begin{align*}
    \hat{f}_S &= \text{MLP}(\text{Norm}(f_S \odot R_{Q,S})) \\
    \hat{f}_Q &= \text{MLP}(\text{Norm}(f_Q \odot R_{Q,S}))
\end{align*}
\]

(6)

where \(\odot\) denotes an element-wise multiplication operation, and CSRMs are defined as a soft relation mask that can strengthen the
features $f$ extracted by CNNs-based backbone. Besides, $MLP$ is the Feed Forward module as illustrated in Fig. 2, which allows the backbone to focus on the subtle feature differences of the predicted semantically similar regions. The experimental analyses of the REP are shown in Sec. 4.4. Overall, $\hat{f}_S$ and $\hat{f}_Q$ are defined as the output features of the REP, and then will be fed into the classification head as illustrated in Fig. 2.

**Differences among Transformer-based Cross-attention Variants.** Given high-level feature pairs ($f_S, f_Q$) extracted from CNN-based backbone $F$, there are many Transformer-based alternatives to model support-query patch-level relations of the extracted high-level features pairs ($f_S, f_Q$). They can be categorized into three classes as follows.

1) Unidirectional cross-attention structure: As shown in Figs. 3(a) and 3(b), only the features from a single branch (support or query branch) are enhanced by means of cross-attention from another branch. For such a case, enhanced features $\tilde{f}$ and original backbone features $f$ are used as the input of classification head. Such a unidirectional cross-attention way fails to achieve high classification accuracy, since this way only considers the semantic enhancement of only a single branch.

2) Bidirectional but asymmetric structure: As illustrated in Figs. 3(c) and 3(d), $S \rightarrow Q$ (or $Q \rightarrow S$) is a sequential stack of the above $S \rightarrow Q$ and $Q \rightarrow S$ variants; $Q \rightarrow S$: Bidirectional asymmetric cross-attention by stacking the $Q \rightarrow S$ and $S \rightarrow Q$ variants.

### 3.3 Overall Objectives and Cross-attention Learning Strategy

**Overall Objectives.** For finding a many-to-many semantic relation matching between the support and query branches, we utilize the multi-head attention mechanism, which consists of multiple attention layers with different token embedding parameters. The overall loss function of the $n$-th class on base classes $D_{base}$ of HelixFormer can be written as follows:

$$
L_{cls}(F, W, \phi, H; x_S) = \sum_{n \in N} \sum_{y_Q \neq y_Q} \log P(y_Q = n|x_Q; x_S) = \frac{\exp(H([\tilde{f}_S, \tilde{f}_Q]))}{\sum_{n' \in N} \exp(H([\tilde{f}_{n'}, \tilde{f}_{n'}]))}
$$

where $W$ and $\phi$ denote learnable parameters in RMP and REP, respectively.

**Cross-attention Learning Strategy.** To transfer the semantic relations from base classes $D_{base}$ to novel classes $D_{novel}$, we employ a two-stage training strategy. Firstly, to ensure that the backbone $F$ has sufficient high-level semantic knowledge for subsequent cross-attention matching process, the $F$ is trained on base classes $D_{base}$ by optimizing Eq. 1. Secondly, we insert the HelixFormer at the end of backbone $F$, and finetune the entire framework to compare the support and query images by optimizing Eq. 7 on $D_{base}$.

### 4 EXPERIMENTS

We evaluate the proposed method on five FSFGL benchmarks including Stanford Dogs, Stanford Cars, NABirds, CUB, and Aircraft. Additionally, we also perform cross-domain few-shot experiments to further show the transferability and adaptability of the HelixFormer. The following experiments are implemented by Pytorch, and all images are resized to 84×84 pixels for fair comparison.

#### 4.1 Dataset

- **Stanford Dogs** [22] contains a total of 20580 images and 120 sub-classes of dogs. Following [71], we adopt 70, 20, 30 classes for meta-train, meta-validation and meta-test, respectively.
- **Stanford Cars** [24] consists of 16,185 images from 196 sub-classes, and we
adopt 130, 17, 49 classes for meta-train, meta-validation and meta-test, respectively. NABirds [47] provides 555 sub-classes of birds from North America. We use 350, 66, 139 categories for meta-training, meta-validation and meta-test, respectively, which is consistent with [19]. CUB [49] has 11,788 bird images containing 200 classes. We follow the commonly used split way [46, 57], which employs 100, 50, and 50 classes for meta-training, meta-validation and meta-test, respectively. Aircraft [34] includes 100 classes and 10000 aircraft images. Following the split way in [57], we use 50, 25, and 25 classes for meta-training, meta-validation and meta-test, respectively.

### 4.2 Experimental Setup

**Stage One: Pre-training Backbone.** Following the setting in [5, 38, 41, 64], we insert a fully-connected layer at the end of the selected backbone such as Conv-4 or ResNet-12, and train the backbone on base classes $D_{base}$. In this stage, the backbone is trained from scratch using SGD optimizer with a batch size of 128, a momentum of 0.9, a weight decay of 0.0005, and an initial learning rate of 0.1. To keep consistent with the setting in [58], the learning rate decays at 85 and 170 epochs. We remove the fully-connected layer for performing the next meta-training stage.

**Stage Two: Meta-training HelixFormer.** In this stage, we first insert the proposed HelixFormer at the end of the pre-trained backbone, and then finetune the whole model to perform cross-attention for each input image pair. A learning rate of 0.001 is employed for all modules. SGD with a weight decay of 0.001 and Adam are used to optimize the backbone and HelixFormer, respectively. The whole training process lasts 130 epochs, and the learning rate decays to 0.0001 and 0.00001 at 70 and 110 epochs, respectively. The number of multi-head attention in the single-layer HelixFormer is set to 2, and the corresponding experimental analysis is shown in Sec. 4.4. For model evaluation, we report the results with 95% confidence intervals over 2000 test episodes, and the best model is chosen according to its accuracy on the validation set.

### 4.3 Experimental Results

**Few-shot Fine-grained Image Classification Results.** It is generally recognized that spatially local feature relations across different fine-grained objects are particularly important for fine-grained classification. Thus, we first validate the effectiveness of HelixFormer on a wide range of fine-grained benchmark datasets, which is shown in Tables 1-3.

| Method                  | Setting | Backbone | 1-shot  | 5-shot  |
|-------------------------|---------|----------|---------|---------|
| RelationNet (CVPR-20) [43] | In.     | Conv-4   | 43.29±0.46* | 55.15±0.39* |
| GNN† (ICLR-18) [12]      | In.     | Conv-4   | 46.98±0.98 | 62.27±0.95 |
| CovaMNet (AAAI-19) [28]  | In.     | Conv-4   | 49.10±0.76 | 63.04±0.65 |
| DN4 (CVPR-19) [27]       | In.     | Conv-4   | 45.73±0.76 | 66.33±0.66 |
| LRPABN (TMM-20) [19]     | In.     | Conv-4   | 45.72±0.75 | 60.94±0.66 |
| MattML (IJCAI-20) [71]   | In.     | Conv-4   | 54.84±0.53 | 71.34±0.38 |
| ATL-Net (IJCAI-20) [9]   | In.     | Conv-4   | 54.92±0.49 | 73.20±0.69 |
| FRN (CVPR-21) [57]       | In.     | Conv-4   | 49.37±0.20 | 67.13±0.17 |
| LSC+SSM (ACM MM-21) [58] | In.     | Conv-4   | 55.53±0.45 | 71.68±0.36 |
| Ours                    | In.     | Conv-4   | 59.81±0.50 | 73.40±0.36 |
| LSC+SSM (ACM MM-21) [58] | In.     | ResNet-12 | 64.15±0.49 | 78.28±0.32 |
| Ours                    | In.     | ResNet-12 | 65.92±0.49 | 80.65±0.36 |

**Table 2: 5-way classification accuracy (%) on the CUB (using bounding-box cropped images).**

| Method                  | Backbone | CUB 1-shot | CUB 5-shot |
|-------------------------|----------|------------|------------|
| RelationNet* (CVPR-18) [43] | Conv-4  | 68.87±0.22 | 82.90±0.15 |
| DeepEMD (CVPR-20) [69] | ResNet-12 | 75.65±0.83 | 88.69±0.50 |
| ICI (CVPR-20) [53]     | ResNet-12 | 76.16       | 90.32      |
| CTX (NIPS-20) [8]      | ResNet-12 | 78.47       | 90.90      |
| FRN (Baseline) [57]    | ResNet-12 | 80.80±0.20  |           |
| FRN (CVPR-21) [57]     | ResNet-12 | 83.16       | 92.59      |
| LSC+SSM (ACM MM-21) [58] | ResNet-12 | 77.77±0.44  | 89.87±0.24 |
| Ours (Baseline)        | ResNet-12 | 72.61±0.47  | 85.60±0.29 |
| Ours                   | ResNet-12 | 81.66±0.30  | 91.83±0.17 |

**Table 3: 5-way classification accuracy (%) on the Aircraft dataset.**
First, Table 1 reports the classification accuracies on the Standard Dogs, Standard Cars, and NABirds. It can be seen from Table 1 that the proposed HelixFormer can be applied on different backbones such as Conv-4 [43] and ResNet-12 [4]. Moreover, we compare the proposed HelixFormer with state-of-the-art general few-shot learning methods (including DN4 [27] and FRN [57], etc.) and few-shot fine-grained image classification methods (including MattML [71], LSC+SSM [58], etc.). FRN considers the few-shot learning as a feature reconstruction problem by reducing the reconstruction errors of images belonging to the same classes, while LSC+SSM attempts to align spatial distribution of feature maps between support and query images, both of which are state-of-the-art general and fine-grained few-shot classification methods, respectively. The experimental results show that the proposed HelixFormer outperforms these methods with a considerable margin, further demonstrating that learning the cross-attention using the proposed double-helix structure can boost the accuracy of fine-grained object recognition.

Furthermore, we also conduct experiments on two more challenging datasets (CUB and Aircraft), as shown in Table 2 and 3. We also observe a consistent accuracy increase using our method. Besides, we would like to emphasize that the HelixFormer significantly boosts the accuracy of Baseline from 72.61% to 81.66% on CUB dataset and has been verified on five commonly used fine-grained datasets, comprehensively showing the effectiveness of HelixFormer on fine-grained recognition.

### Few-shot Fine-grained Cross-domain Image Classification Results

Considering that the distribution differences between the training and test data often exist, we conduct few-shot fine-grained recognition experiments under cross-domain setting, to validate the effectiveness of the HelixFormer in alleviating the impact of the domain differences, and the results are reported in Table 4.

We carry out the cross-domain adaptation from generic bird categories (widely collected from Internet) to a particular country (America). It can be seen from Table 4 that our method has higher accuracy than both the Baseline model and LSC+SSM model [58], demonstrating that HelixFormer also can improve the transferability and domain adaptability of the existing models.

### 4.4 Insight Analyses

#### Results of using Fully-connected or Convolutional-based Token Embedding
The choice on feature embedding way (in a fully-connected or convolutional way) of the input tokens is essential to guarantee the good performance of the proposed HelixFormer.
Figure 5: Visualization results of features extracted by backbone, the RMP, and the REP, respectively. Due to the global feature variations (e.g., the color of an object) and local feature changes (e.g., headlight translation of a car and beak rotation of a bird), it is hard to find the cross-image patch-level matching of semantic features, as shown in the heatmaps from the backbone. By HelixFormer, the key cross-image object semantic relations, such as birds’ wings or cars’ headlights, can be effectively matched. 

Please refer to our supplementary material for more visualization results.

| Method          | SY? | CUB 1-shot | CUB 5-shot | Stanford Cars 1-shot | Stanford Cars 5-shot |
|-----------------|-----|------------|------------|----------------------|----------------------|
| RN [43]         | -   | 66.08±0.50 | 79.04±0.35 | 56.55±0.52           | 70.52±0.39           |
| RN with S→Q     | X   | 77.66±0.48 | 89.39±0.27 | 74.50±0.50           | 89.89±0.22           |
| RN with Q→S     | X   | 77.46±0.47 | 89.86±0.25 | 74.42±0.50           | 88.63±0.24           |
| RN with S→Q w/o | ✓   | 79.34±0.45 | 91.01±0.24 | 75.46±0.37           | 89.68±0.25           |

Table 7: The study of bidirectional cross-attention using asymmetric or symmetric structure. SY is short for symmetric, and S→Q denotes the asymmetric but bidirectional structure and the definition of S→Q follows Table 5.

| Method          | REP? | CUB 1-shot | CUB 5-shot | Stanford Cars 1-shot | Stanford Cars 5-shot |
|-----------------|------|------------|------------|----------------------|----------------------|
| RN [43]         | -    | 66.08±0.50 | 79.04±0.35 | 56.55±0.52           | 70.52±0.39           |
| RN with S→Q     | w/o  | 77.90±0.45 | 89.39±0.27 | 73.52±0.49           | 88.48±0.25           |
| RN with S→Q     | with | 79.34±0.45 | 91.01±0.24 | 75.46±0.37           | 89.68±0.25           |

Table 8: The results of removing the REP in HelixFormer.

| Method          |  # Multi-head | CUB 1-shot | CUB 5-shot | Stanford Cars 1-shot | Stanford Cars 5-shot |
|-----------------|---------------|------------|------------|----------------------|----------------------|
| RN with S→Q     | 1             | 77.52±0.49 | 74.62±0.50 |                      |                      |
| RN with S→Q     | 2             | 79.34±0.45 | 75.46±0.37 |                      |                      |
| RN with S→Q     | 4             | 78.68±0.47 | 73.62±0.48 |                      |                      |

Table 9: Results of changing the number of multi-head attention.

Asymmetric and Symmetric RMP. For bidirectional attention, we observe from Table 7 that a symmetrical structure has an advantage in improving the model accuracy. We further visualize their heatmaps in Fig. 4, and find that the symmetrical structure captures relatively more accurate inter-object semantic relations.

The Role of REP. We further show the effectiveness of the REP in the proposed HelixFormer, by directly feeding the CSRMs pairs (R_Q S, R_S Q) as the input of the classification head. We observe the performance drop by comparing the last two rows in Table 8, showing the importance for distinguishing subtle feature differences.

Multi-head Attention in HelixFormer. Table 9 shows the 5-way 1-shot classification accuracy by changing the number of multi-head attention using Conv-4 backbone.

Study of Parameter Size and Feature Visualization. In the few-shot learning community, models with smaller parameter sizes are usually adopted to avoid over-fitting for the limited number of training samples. In other words, increasing the number of model parameters may not improve its generalization performance. Table 10 shows that the HelixFormer more effectively boosts the model accuracy, compared with other models with more parameters and FLOPs. Furthermore, we visualize the support and query features which are extracted from the backbone, the proposed RMP, and REP of HelixFormer respectively. The visualized results are illustrated in Fig. 5, and the heatmaps are obtained via a max-pooling operation along the channel dimension of feature maps to preserve the spatial information. Besides, we also visualize the support-query features using asymmetric or symmetric cross-attention structure in Fig. 5. These visualization results illustrate that the semantically similar regions between support and query images can be well matched via HelixFormer.

5 CONCLUSION

In this work, to exploit the semantic relations across different fine-grained objects, we proposed a Transformer-based double-helix cross-attention model, namely HelixFormer, which is composed of a Relation Mining Process (RMP) to discover the cross-image object semantic relation and produce patch-level cross-relation maps, and a Representation Enhancement Process (REP) to enhance the identified discriminative local features for final recognition. Experimental results demonstrate that such a bidirectional and symmetrical structure has the merit of ensuring the cross-image object semantic relation consistency, thus further improving the model generalization ability on few-shot fine-grained learning.

ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China (No. 62071127 and No. 62101137), Zhejiang Lab Project (No. 2021KH0AB05).
[59] Ziyang Wu, Yuwei Li, Lihua Guo, and Kui Jia. 2019. Parn: Position-aware relation networks for few-shot learning. In Proceedings of the IEEE/CVF international conference on computer vision. 6659–6667.

[60] Kaicheng Yang, Hua Xu, and Kai Gao. 2020. CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis. In Proceedings of the 28th ACM International Conference on Multimedia. 521–528.

[61] Shuo Yang, Lu Liu, and Min Xu. 2021. Free Lunch for Few-shot Learning: Distribution Calibration. In ICLR.

[62] Yukuan Yang, Fangyu Wei, Miaoqing Shi, and Guoqi Li. 2020. Restoring negative information in few-shot object detection. In NeurIPS.

[63] Chuangguan Ye, Hongyuan Zhu, Yongbin Liao, Yanggang Zhang, Tao Chen, and Jiayuan Fan. 2022. What Makes for Effective Few-Shot Point Cloud Classification? In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 1829–1838.

[64] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. 2020. Few-shot learning via embedding adaptation with set-to-set functions. In CVPR.

[65] Mingzhang Yin, George Tucker, Mingyuan Zhou, Sergey Levine, and Chelsea Finn. 2020. Meta-learning without memorization. In ICLR.

[66] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. 2021. HRFormer: High-Resolution Transformer for Dense Prediction. In NeurIPS.

[67] Bo Zhang, Tao Chen, Bin Wang, and Runyao Li. 2021. Joint Distribution Alignment via Adversarial Learning for Domain Adaptive Object Detection. IEEE Transactions on Multimedia (2021).

[68] Bo Zhang, Hancheng Ye, Gang Yu, Bin Wang, Yike Wu, Jiayuan Fan, and Tao Chen. 2022. Sample-Centric Feature Generation for Semi-Supervised Few-Shot Learning. IEEE Transactions on Image Processing (2022).

[69] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. 2020. DeepEMD: Few-Shot Image Classification With Differentiable Earth Mover’s Distance and Structured Classifiers. In CVPR.

[70] Ruixiang Zhang, Tong. Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. 2018. Metagan: An adversarial approach to few-shot learning. In NeurIPS.

[71] Yashui Zhu, Chenlong Liu, and Shuqiang Jiang. 2020. Multi-attention meta learning for few-shot fine-grained image recognition. In IJCAI.

[72] Feiqin Zhuang, Yali Wang, and Yu Qiao. 2020. Learning attentive pairwise interaction for fine-grained classification. In AAAI.