Detect Any Keypoints: An Efficient Light-Weight Few-Shot Keypoint Detector

Changsheng Lu[1], **Piotr Koniusz**[2,1]

[1]The Australian National University
[2]Data61/CSIRO
changshengluu@gmail.com, piotr.koniusz@data61.csiro.au

Abstract

Recently the prompt-based models have become popular across various language and vision tasks. Following that trend, we perform few-shot keypoint detection (FSKD) by detecting any keypoints in a query image, given the prompts formed by support images and keypoints. FSKD can be applied to detecting keypoints and poses of diverse animal species. In order to maintain flexibility of detecting varying number of keypoints, existing FSKD approaches modulate query feature map per support keypoint, then detect the corresponding keypoint from each modulated feature via a detection head. Such a separation of modulation-detection makes model heavy and slow when the number of keypoints increases. To overcome this issue, we design a novel light-weight detector which combines modulation and detection into one step, with the goal of reducing the computational cost without the drop of performance. Moreover, to bridge the large domain shift of keypoints between seen and unseen species, we further improve our model with mean feature based contrastive learning to align keypoint distributions, resulting in better keypoint representations for FSKD. Compared to the state of the art, our light-weight detector reduces the number of parameters by 50%, training/test time by 50%, and achieves 5.62% accuracy gain on 1-shot novel keypoint detection in the Animal pose dataset. Our model is also robust to the number of keypoints and saves memory when evaluating a large number of keypoints (*e.g.*, 1000) per episode.

Introduction

Keypoint detection is a critical research topic in computer vision. As it can provide concise semantic and structural information, it has many applications in pose estimation of humans (Cao et al. 2019b) and animals (Pereira et al. 2019), behavior analysis (Graving et al. 2019), face alignment (Kowalski, Naruniec, and Trzcinski 2017), and fine-grained image classification (Tang, Wertheimer, and Hariharan 2020), *etc*. Over the past decade, the deep keypoint detection evolved significantly. However, the fully-supervised keypoint detectors are limited by the training data, and recognize only specific body parts and species, failing to generalize to unseen species. Semi-supervised methods suffer from similar issues and require several hundreds or thousands of labels for training on new class, and unsupervised methods fail to detect keypoints desired by users. Moreover, due to the diversity of animal species, one cannot collect data samples of all species, which justifies research on few-shot keypoint detection (FSKD): *having trained the model on a diverse dataset, one can use some similarity measure to rapidly recognize novel/base keypoints in unseen species given only one or a few labeled samples*.

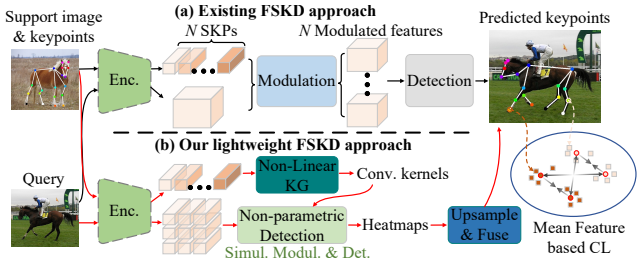


Figure 1: Comparison of FSKD methods. Existing methods perform feature modulation per support keypoint prototype (SKP) followed by a detection step. In contrast, our model uses conditionally generated kernels and a non-parametric detection module to conduct simultaneous modulation and detection. Our model avoids generating the expensive "modulated" feature maps, thus significantly reducing the number of parameters, memory usage & computational time.

Recent works on detecting keypoints from few samples can be broadly divided into two categories. The first category, class-specific few-shot keypoint detection (He et al. 2023; Yao et al. 2021), uses a large number of unlabeled datapoints and few labeled samples to perform landmark localization. Such a family of methods works with specific keypoint types but is unable to handle varying number of keypoints over diverse unseen species.

In contrast, the second category is based on the general few-shot keypoint detection (Lu and Koniusz 2022; Bohdal et al. 2023; Lu, Zhu, and Koniusz 2023) inspired by few-shot learning (Sung et al. 2018; Koch et al. 2015). Such FSKD methods use episodes for training and evaluation. Our work falls into this category as it offers the flexibility of detecting any keypoints in a query image given the prompts of support keypoints. However, there exist two important issues that hinder FSKD, which are i) *the scalability issue w.r.t. the number of keypoints per episode* and ii) *the large domain*

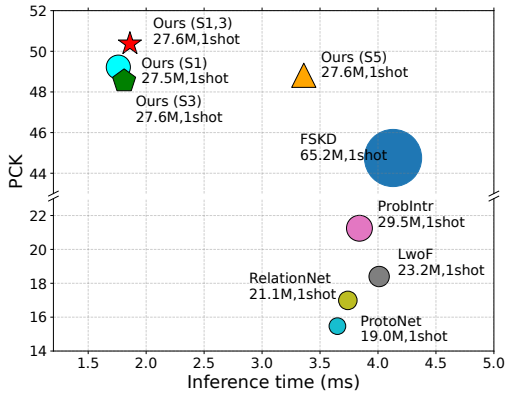


Figure 2: The 1-shot novel keypoint detection (the Animal pose dataset). Compared to the previous best FSKD model, ours enjoys lower inference time and higher performance.

shift of keypoints between seen and unseen types of species.

Fig. 1a illustrates existing FSKD methods which mainly follow the modulation-detection separate (MDS) design. Each support keypoint prototype (SKP) is modulated with a query feature map via attention, yielding modulated feature map for subsequent keypoint detection. Such a design is similar to few-shot object detection (FSOD) (Fan et al. 2020; Kang et al. 2019; Zhang et al. 2020). Though impressive, it is not robust w.r.t. the number of keypoints as each prompted keypoint would result in a modulated feature map. It would lead out-of-memory and slow speed if large number of keypoints are prompted. To solve this issue, we propose to refine SKP into better representations: we form one or multiple groups of convolutional kernels with diverse spatial resolutions for performing simultaneous modulation and detection (SMD). In contrast to existing methods (Bohdal et al. 2023; Lu and Koniusz 2022), our method directly outputs a keypoint heatmap after a non-parametric detection, avoiding the byproduct of intermediate modulated feature maps, thus saving memory and computation time. Our method is simple and it does not sacrifice its performance. Based on our theoretical analysis and experimental results, we argue that the SMD-based model enjoys equal or better feature modeling capability compared to MDS-based models.

When detecting keypoints in unseen species, the FSKD model faces great challenges as the seen and unseen species significantly differ in their appearance, pose, keypoint types and numbers. Thus, we propose an efficient way of narrowing the domain shift and improving the generalization of our lightweight model. As we have access to some base species in the training phase, we propose to align representations of the same keypoint type across different species to reduce the domain shift, while repelling representations of different keypoint types. As an example, we align representations of "nose" across several species, *e.g.*, cat, dog, cow, while repelling representations of "nose" and "eye". To this end, inspired by contrastive learning (CL) (Chen et al. 2020; Wu et al. 2018), we propose mean feature based contrastive learning (MFCL) to improve our model to learn better keypoint features in deep metric space. Our MFCL enjoys better results than instance-based CL for FSKD as the mean features improve the stability of keypoint representations. Moreover, as negative samples have impact on CL, we propose to control the hardness of negative keypoints sampled from image, which improves the diversity of negatives.

In summary, the contributions of this paper are: 1) We propose a lightweight few-shot keypoint detector which scales gracefully w.r.t. the number of keypoints. Despite simplicity, our model performs well on detection of keypoints on unseen species; 2) We further improve our lightweight detector with mean feature based contrastive learning (MFCL) and a novel method of negative keypoints control. MFCL improves detection scores on both base and novel keypoints; 3) Our model reduces computational cost while achieving the new state-of-the-art performance on various datasets such as the popular Animal pose dataset (Cao et al. 2019a), CUB (Wah et al. 2011), NABird (Van Horn et al. 2015), and AwA (Banik, Li, and Dong 2021).

Related Work

Keypoint Detection. As a long-standing pursuit, keypoint detection has been extensively studied in literature, ranging from the traditional interest point (Lowe 2004; Derpanis 2004) to deep corner detection (Zhao et al. 2023), semi-supervised (Moskvyak et al. 2021) and fully-supervised keypoint estimation (Tompson et al. 2014; Cao et al. 2019b; Cheng et al. 2020; Fang et al. 2017; Sun et al. 2019). Current deep keypoint localization methods follow two paradigms: i) direct regression on keypoint coordinates (Carreira et al. 2016), and ii) heatmap regression followed by coordinate decoding (Cheng et al. 2020), which usually leads to higher performance due to its simplicity of spatial mapping. We also adopt heatmap regression but unlike existing heatmap-based approaches dedicated to specific body parts, *e.g.*, top-down (Sun et al. 2019; Fang et al. 2017; He et al. 2017) and bottom-up pose estimators (Newell, Yang, and Deng 2016; Cao et al. 2019b; Cheng et al. 2020), our few-shot model offers more flexible keypoint detection, breaking the limitation of keypoint types to be detected.

Few-shot Keypoint Detection. Compared to the widely studied supervised keypoint detection, the emerging FSKD is still under-explored. Since general FSKD shares close relationship with few-shot learning (FSL), many well-known FSL methods such as ProtoNet (Snell, Swersky, and Zemel 2017), RelationNet (Sung et al. 2018), LwoF (Gidaris and Komodakis 2018) and MAML (Finn, Abbeel, and Levine 2017) have been extended to the problem of keypoint detection. Both LwoF (Gidaris and Komodakis 2018) and our work generate parameters, but differ in three aspects: 1) our kernel generator (KG) generates diverse resolutions for kernels; 2) our KG is non-linear and we present a theoretical analysis that SMD has potential not to sacrifice performance while saving computational time; 3) we apply non-linear KG in FSKD. Compared to other FSL tasks (Shi et al. 2023b; Fan et al. 2020), keypoints are smaller and harder to detect than objects, thus it requires more stable representations and local & global contexts to perform robust detection.

Recently, Ge, Zhang, and Luo (2021) applied FSKD for the fashion landmark detection, which reduced the expensive annotations on unseen clothes. Bohdal et al. (2023) pro-

pose a dataset-of-datasets, benchmarking the FSL algorithm universal to various vision tasks. Lu and Koniusz (2022) formalize four comprehensive FSKD settings by the combinatorial configuration of detecting base or novel keypoints on seen or unseen species during testing phase. Moreover, the localization uncertainty is modeled to alleviate the influence of local noise. FSKD can also benefit pose estimation, *e.g.*, class-agnostic pose estimation in animals (Lu and Koniusz 2022; Xu et al. 2022; Shi et al. 2023a). Compared to pioneering work (Lu and Koniusz 2022), this paper aims to address the issue of scalability w.r.t. the number of keypoints, reduce model parameters, and leverage domain distribution alignment to improve keypoint detection on unseen species.

Contrastive Learning for Few-shot Model. Contrastive learning (CL) is popular in self-supervised visual representation learning (Chen et al. 2020; Wu et al. 2018; Bardes, Ponce, and LeCun 2021; He et al. 2020; Grill et al. 2020). It exploits internal data structure via *instance discrimination*. The core idea is to cluster positive sample pairs while repelling negative samples. Recently, a variety of works (Yang, Wang, and Zhu 2022; Liu et al. 2021; Ouali, Hudelot, and Tami 2021; Doersch, Gupta, and Zisserman 2020) have explored how to improve few-shot models via contrastive learning. Both our work and (Yang, Wang, and Zhu 2022; Liu et al. 2021) use labels given by training episodes to perform supervised contrastive learning (Jian, Gao, and Vosoughi 2022), whose objective loss enjoys a generalization of triplet (Weinberger and Saul 2009) and N-pair losses (Sohn 2016). However, in contrast to improving few-shot model via instance-based CL, we use mean feature based contrastive learning (MFCL), which estimates the mean feature of keypoints at the episode level and we perform contrastive learning between episodes. We observe that MFCL yields consistent improvements compared to instance-based CL for FSKD.

Methodology

Problem Definition

Let us denote the set of training species as $\mathcal{C} = \{c_i\}_{i=1,2,\cdots,N_\mathcal{C}}$ and the set of testing species as $\mathcal{C}' = \{c'_i\}_{i=1,2,\cdots,N_{\mathcal{C}'}}$, where each element represents one class of species. Let the set of training keypoint types be $\mathcal{X} = \{k_i\}_{i=1,2,\cdots,N_\mathcal{X}}$ and the testing keypoint types be $\mathcal{X}' = \{k'_i\}_{i=1,2,\cdots,N_{\mathcal{X}'}}$.

Then, FSKD needs to learn a model on the training species and keypoints $\mathcal{C} \times \mathcal{X}$ and generalize it to test ones $\mathcal{C}' \times \mathcal{X}'$, where the hardest setting is $\mathcal{C} \cap \mathcal{C}' = \varnothing, \mathcal{X} \cap \mathcal{X}' = \varnothing$. Following general few-shot learning (Vinyals et al. 2016; Sung et al. 2018), FSKD model is evaluated on episodes, each of which includes sampled query image and support image with keypoint annotations. FSKD aims to detect the corresponding keypoints in query image. If there are N support keypoints and K support images, the problem is defined as N-way K-shot detection.

Lightweight Few-shot Keypoint Detection

The proposed lightweight FSKD model consists of four steps: i) feature extraction, ii) keypoint feature aggregation, iii) simultaneous modulation and detection (SMD), and iv) heatmap upsampling and fusion. Despite simplicity, each meticulously designed module makes the whole composition very efficient yet highly performant.

Keypoint Feature Aggregation The FSKD model takes episode as input. The support and query images $\mathbf{I}^\text{s}$ and $\mathbf{I}^\text{q} \in \mathbb{R}^{3\times l_0 \times l_0}$ are encoded as support and query feature maps $\mathcal{F}(\mathbf{I}^\text{s})$ and $\mathcal{F}(\mathbf{I}^\text{q})$ in feature space $\mathbb{R}^{C\times l\times l}$ via a weight-shared encoder $\mathcal{F}$. The encoded feature map should be extracted from high-level layers as they contain richer semantics and context that help infer keypoint locations. Subsequently, a keypoint feature aggregator $\mathcal{A}$ is employed to aggregate the local features for each support keypoint. We use pixel-wise weighted summation between Gaussian heatmap and support feature map $\mathcal{F}(\mathbf{I}^\text{s})$ to build support keypoint representation (SKR) as

$$\mathbf{\Phi}_{k,n} = \mathcal{A}(\mathcal{F}(\mathbf{I}^\text{s}), \mathbf{x}_{k,n}; \sigma) \tag{1}$$

where $\mathbf{\Phi}_{k,n} \in \mathbb{R}^C$ is the SKR aggregated at $\mathbf{x}_{k,n}$ for n-th keypoint in k-th support image; $\sigma = \frac{l}{l_0}\sigma_0$ is the standard deviation that controls Gaussian spread.

Simultaneous Modulation and Detection Once the support keypoint representations (SKR) are obtained, they are subsequently correlated with a query feature map to guide the network to detect query keypoints.

The major bottleneck in few-shot learning is the limited number of support samples, which means the extracted support features are non-representative of the true class distribution. To mitigate this issue, several works (Snell, Swersky, and Zemel 2017; Sung et al. 2018) use the mean estimator to build a class prototype, and then use the class prototype for few-shot image classification. This simple yet effective approach is adopted by many few-shot keypoint detection (FSKD) works (Bohdal et al. 2023; Lu and Koniusz 2022). For example, by averaging SKRs $\mathbf{\Phi}_{k,n}$ across K support images, one can build support keypoint prototype (SKP) as $\mathbf{\Psi}_n = \frac{1}{K}\sum_k \mathbf{\Phi}_{k,n}$ for support-query modulation and detection. However, we argue that this may not be optimal approach for FSKD. As the number of SKPs is limited in one-shot case and they may be noisy due to noisy support samples, they may affect the subsequent detection step. However, the SKP can be refined via a non-linear module to yield better representations for modulation (see Fig. 1). This motivates us to design a non-linear kernel generator (KG), which is conditioned on SKP to generate convolutional kernels for the detection step. Moreover, we discover that the coupled non-linear KG and the non-parametric detection step can realize simultaneous modulation and detection, solving scalability issue w.r.t. the number of keypoints in FSKD.

Considering both efficiency and lightweight model requirements, we devise the non-linear kernel generator (KG) as *space-channel disentangled* refinement network

$$\mathbf{\Pi} = (\mathbf{\Pi}_\text{sp}^{G_1}, \cdots, \mathbf{\Pi}_\text{sp}^{G_S}) \circ \mathbf{\Pi}_\text{ch} \tag{2}$$

where $\mathbf{\Pi}_\text{sp}^{G_S}$ is the *space refinement module* that expands the resolution of SKPs $\mathbf{\Psi} \in \mathbb{R}^{N\times C}$ into kernels $\mathbf{W}^{G_S} \in$

$\mathbb{R}^{N\times C\times S\times S}$ at resolution S, while $\mathbf{\Pi}_{\text{ch}}$ is the *channel refinement module* that refines channels of $\mathbf{W}^{G_S}$, which is shared across space refinement modules. Via the collaboration of space and channel refinements, our kernel generator is able to generate multi-group kernels

$$\mathbf{W}^{G_1} = \mathbf{\Pi}_{\text{ch}}(\mathbf{\Pi}^{G_1}_{\text{sp}}(\mathbf{\Psi})),\cdots,\mathbf{W}^{G_S} = \mathbf{\Pi}_{\text{ch}}(\mathbf{\Pi}^{G_S}_{\text{sp}}(\mathbf{\Psi})) \quad (3)$$

with different resolutions $1,\cdots,S$. The generation of diverse kernel resolutions improves the *correlation window* during modulation and detection. $\mathbf{\Pi}^{G_S}_{\text{sp}}$ can be easily realized by a depth-wise deconvolution, and $\mathbf{\Pi}_{\text{ch}}$ by a non-linear block with 1×1 conv. We provide the detailed architecture of the kernel generator (KG) in **Suppl. Mat.**[1]

Subsequently, we inject the multi-group kernels $\mathbf{W}^* = \{\mathbf{W}^{G_1},\cdots,\mathbf{W}^{G_S}\}$ into a non-parametric detection module $\mathcal{D}$ to perform simultaneous modulation and detection. Specifically, as each group $\mathbf{W}^{G_S}$ has N kernels, each kernel $\mathbf{W}^{G_S}_n$ is of size $C\times S\times S$. We use $\mathbf{W}^{G_S}_n$ as a filter to perform *scaled convolution* with the query feature map $\mathcal{F}(\mathbf{I}^{\text{q}})\in\mathbb{R}^{C\times l\times l}$, yielding the heatmap as

$$\mathbf{H}^{G_s}_n = \frac{1}{S\times S\times\sqrt{C}}\left(\mathcal{F}(\mathbf{I}^{\text{q}})\otimes\mathbf{W}^{G_S}_n\right), \quad (4)$$

where $\mathbf{H}^{G_s}_n\in\mathbb{R}^{l\times l}$ is an n-th single-channel heatmap. Consequently, each group of kernels $\mathbf{W}^{G_S}$ result in N heatmaps $\mathbf{H}^{G_s}\in\mathbb{R}^{N\times l\times l}$, corresponding to N support keypoints.

Discussion. Despite the simplicity of the proposed non-linear kernel generator $\mathbf{\Pi}$ and non-parametric detection module $\mathcal{D}$, we emphasize that their combination prevents the use of intermediate modulation features as in (Bohdal et al. 2023; Lu and Koniusz 2022). Our method directly generates heatmaps via simultaneous modulation and detection (SMD), thus solving the scalability issue w.r.t. the number of keypoints. Our method also provides an opportunity to refine the SKP into better representations for $\mathcal{D}$. We present a theoretical analysis that our SMD design can achieve fitting capability (Yosida 2012) equal/greater than modulation-detection separate (MDS) design.

Proposition 1 *Let $\mathbf{X}\in\mathbb{R}^{C\times H\times W}$ be query feature map and $\mathbf{a}\in\mathbb{R}^C$ be a single SKP. For the MDS design, consider a two linear-layer detection head $f(\mathbf{Z}) = \mathbf{W}_{1,2}\phi(\mathbf{W}_{1,1}\mathbf{Z})$, where $\mathbf{W}_{1,1}\in\mathbb{R}^{C\times C}$ and $\mathbf{W}_{1,2}\in\mathbb{R}^{1\times C}$ are the weights of f, ϕ is an activation function,* i.e.*, softplus, and $\mathbf{Z}\in\mathbb{R}^{C\times H\times W}$ is the modulated feature map via $\mathbf{Z} = \mathbf{X}\odot\mathbf{a}$, the $\odot$ denotes the channel-wise multiplication. Let $\mathbf{H}_f = f(\mathbf{Z})$ be the output heatmap. For our SMD design, let kernel generator be $g(\mathbf{a}) = \mathbf{W}_{2,2}\phi(\mathbf{W}_{2,1}\mathbf{a})$, where $\mathbf{W}_{2,1}$, $\mathbf{W}_{2,2}\in\mathbb{R}^{C\times C}$ are the weights of g. Thus, the output heatmap is $\mathbf{H}_g = \frac{1}{c}g(\mathbf{a})^\intercal\mathbf{X}$, where c is a constant. Both $\mathbf{H}_f$ and $\mathbf{H}_g\in\mathbb{R}^{H\times W}$. If let $\mathbf{W}_{mds} = \mathbf{W}_{1,2}\mathbf{W}_{1,1}$, $\mathbf{W}_{smd} = \mathbf{W}_{2,2}\mathbf{W}_{2,1}$ and expanding ϕ at $\boldsymbol{\varepsilon}\geq 0$ via Taylor expansion, then the difference between $\mathbf{H}_f$ and $\mathbf{H}_g$ can be approximated as*

$$\begin{aligned}\|\mathbf{H}_f-\mathbf{H}_g\|_2 &= \|\mathbf{W}_{1,2}\phi(\mathbf{W}_{1,1}\mathbf{Z}) - \frac{1}{c}g(\mathbf{a})^\intercal\mathbf{X}\|_2\\ &= \|\mathbf{a}^\intercal\big(\text{Diag}(\mathbf{W}_{\text{mds}}) - \frac{1}{c}\mathbf{W}^\intercal_{\text{smd}}\big)\mathbf{X}\|_2.\end{aligned} \quad (5)$$

[1]https://alanlusun.github.io/files/202401-AAAI24-suppl.pdf

Eq. 5 tells that the difference of predicted heatmaps between $\mathbf{H}_f$ and $\mathbf{H}_g$ is zero if $\text{Diag}(\mathbf{W}_{\text{mds}}) = \frac{1}{c}\mathbf{W}^\intercal_{\text{smd}}$ holds. The proposed SMD has fitting capability approximate to the MDS design, which explains that the proposed lightweight FSKD has potential not to lose performance while it can significantly reduce the computational cost.

Heatmap Fusion and Supervision From G_S groups of kernels $\{\mathbf{W}^{G_1},\cdots,\mathbf{W}^{G_S}\}$, one can obtain G_S groups of predicted heatmaps $\{\mathbf{H}^{G_1},\cdots,\mathbf{H}^{G_S}\}$. Each group of heatmaps $\mathbf{H}^{G_S}\in\mathbb{R}^{N\times l\times l}$ is induced by one group of kernels $\mathbf{W}^{G_S}$. For the heatmap regression based keypoint localization, the higher resolution of heatmap can greatly reduce the coordinate encoding and decoding error. Thus, we adopt G_S upsampling modules $\mathcal{U} = (\mathcal{U}^{G_1},\cdots,\mathcal{U}^{G_S})$ to perform group-specific upsampling. Each $\mathcal{U}^{G_S}$ is very lightweight, only including a bilinear upsampler and two *single-channel* 5×5 conv layers. Using single-channel convolution helps further refine heatmaps, *i.e.*, $\mathbf{H}^{G_S} := \mathcal{U}^{G_S}(\mathbf{H}^{G_S})$. During testing phase, we fuse the upsampled multi-group heatmaps as final output $\mathbf{H}\in\mathbb{R}^{N\times ul\times ul}$ (u is an upsampling factor):

$$\mathbf{H} = \frac{1}{G_S}\sum\nolimits_{g=1}^{G_S}\mathbf{H}^g. \quad (6)$$

During the training stage, inspired by HigherHRNet (Cheng et al. 2020), we leverage the multi-group supervision for heatmaps as

$$\mathcal{L}_{\text{hm}} = \frac{1}{G_S}\sum\nolimits_{g=1}^{G_S}\|\mathbf{H}^g-\mathbf{H}^*\|_2^2, \quad (7)$$

where $\mathbf{H}^*$ is the groundtruth heatmap that encodes a query keypoint. In the next section, we explore how to improve our model by introducing a contrastive loss over keypoints.

Improving FSKD by Contrasting Keypoints

The features extracted from few-shot model are notoriously biased to seen classes (Hou et al. 2019), which deteriorates recognition of unseen objects. This occurs because the learned features are less general and transferable. Moreover, due to the limited number of support samples, the extracted representations are hard to represent the novel classes well. Consequently, improving the representation learning ability of our lightweight model should benefit few-shot keypoint detection on unseen species.

To reduce the domain shift of different species while encouraging the similarity of same type of keypoints, one naive solution is to align the keypoint representations across species (s,s') via loss $\mathcal{L}(\mathbf{\Phi}^s,\mathbf{\Phi}^{s'})$. However, a good keypoint representation should be sufficiently discriminative to distinguish it from other patterns such as other irrelevant keypoints or background, while maintaining the similarity between the identical type of keypoints. Thus, contrastive learning (CL) is a good candidate for our model. With CL, we obtain more general and transferable features for FSKD.

Intuitively, one can directly take the already extracted support keypoint representations (SKR) to perform instance level CL. However, it is suboptimal. Since we also have access to query keypoint labels during the training phase, we can leverage all images per episode to build more stable

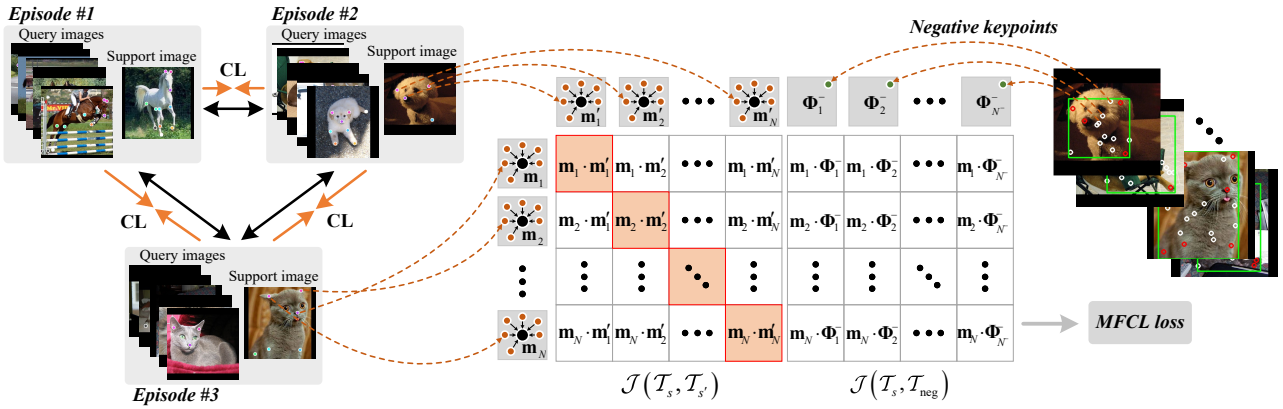


Figure 3: Illustration of mean feature based contrastive learning (MFCL). We show i) episode-level contrastive learning (CL), ii) MFCL between pairwise episodes, and iii) MFCL with negative keypoints, from the left to middle and to the right. In each episode, stable mean features are estimated from same type of keypoints for CL. The negatives are sampled from all images of the pairwise episodes (see rightmost), where the white circles show negative keypoints and the red ones show the anchors.

mean keypoints and perform episode-level CL. This simple idea effectively improves the feature learning and alleviates the noisy SKRs. Let the number of query image be M. Again, we use the keypoint feature aggregator $\mathcal{A}$ to extract the query keypoint representations (QKR). Combine SKRs and QKRs, we obtain a keypoint representation set $\mathcal{T} = \{\mathbf{\Phi}_{k,n}\}_{k=1,n=1}^{K+M,N}$, in total $N(K+M)$ instances per episode. By averaging across all images, we obtain the *mean keypoint feature* $\mathbf{m}_n \in \mathbb{R}^C$ as

$$\mathbf{m}_n = \frac{1}{K+M}\sum\nolimits_{k=1}^{K+M} \mathbf{\Phi}_{k,n}, \tag{8}$$

where $n = 1, \cdots, N$, corresponding to N types of keypoints. It should be noted that $\mathbf{m}_n$ is more representative than SKP $\mathbf{\Psi}_n$. If randomly sampling *two* species (s, s') at a time, each species pertaining to an episode, we have pairwise sets of mean keypoint features, *i.e.*, $\mathcal{T}_s = \{\mathbf{m}_n\}_{n=1}^N$ and $\mathcal{T}_{s'} = \{\mathbf{m}'_n\}_{n=1}^N$, which form a similarity matrix as

$$\mathbf{J}(\mathcal{T}_s, \mathcal{T}_{s'}) = \begin{pmatrix} \cos(\mathbf{m}_1, \mathbf{m}'_1) & \cdots & \cos(\mathbf{m}_1, \mathbf{m}'_N) \\ \vdots & & \vdots \\ \cos(\mathbf{m}_N, \mathbf{m}'_1) & \cdots & \cos(\mathbf{m}_N, \mathbf{m}'_N) \end{pmatrix}, \tag{9}$$

where $\cos(\cdot,\cdot)$ denotes the cosine similarity. Then, the contrastive loss in the direction of species s to s' becomes

$$\mathcal{L}_{\text{CL}}^{s\to s'} = -\langle \mathbb{I}, \log(\text{softmax}(\mathbf{J}(\mathcal{T}_s, \mathcal{T}_{s'})/\tau)) \rangle \tag{10}$$

where $\mathbb{I}$ is the identity matrix and $\langle\cdot,\cdot\rangle$ denotes the inner product. Finally, the overall keypoint contrastive loss $\mathcal{L}_{\text{CL}}$ is

$$\mathcal{L}_{\text{CL}} = \frac{1}{2}(\mathcal{L}_{\text{CL}}^{s\to s'} + \mathcal{L}_{\text{CL}}^{s'\to s}). \tag{11}$$

As the loss $\mathcal{L}_{\text{CL}}$ contrasts *mean keypoint features* instead of *keypoint instances*, and is applied on the *episode level* instead of *image level*, *e.g.*, on a basis of pairwise episodes, we term it as *mean feature based contrastive learning (MFCL)*.

MFCL Extension to Multi-species. Current MFCL only considers contrasting pairwise species (s, s') (corresponding to two episodes). As there are multiple species accessible in training classes $\mathcal{C}$, and one may intend to contrast keypoints across multiple species at each iteration of optimization. Thus, we can drawing multiple episodes at a time and then traverse all pairwise episodes. Then, the contrastive loss for multi-species becomes

$$\mathcal{L}_{\text{CL}}^{N_e} = \sum\nolimits_{s}^{N_e} \sum\nolimits_{s',s'\geq s}^{N_e} \mathcal{L}_{\text{CL}}(s, s') \tag{12}$$

where N_e is the number of episodes drawn per iteration.

Negative Keypoints Control. Contrastive learning benefits from negative keypoints, especially the hard negatives. However, it is difficult to guarantee the arbitrarily sampled point from an image to have "semantic label" different from an anchor point, as point-wise "semantic label" is unavailable or hard to achieve. As a compromise, one can sample points from the foreground or background, but distant to anchors (*i.e.*, support and query keypoints used in MFCL). The anchors with different labels could form hard negative pair, *e.g.*, the ("left ear", "right ear"). We observe there is a high chance that: the foreground points adjacent to anchor may form hard or moderate negatives; the background points distant to anchor may form easy negatives. Thus, we propose to sample the negative points from a scaled bounding box $\mathcal{B}_\alpha = (x_c, y_c, \alpha w, \alpha h)^\intercal$, where (x_c, y_c) is center, (w, h) is width and height, and α is scale factor. Note that $\mathcal{B}_{1.0}$ is the bounding box circumscribed to object and estimated by anchors[1]. When sampling negative points within $\mathcal{B}_\alpha$, we could set the distance threshold to anchors ρ, and the number of negative points N_{neg}. By setting $(\alpha, \rho, N_{\text{neg}})$, we could realize controlling *hardness* and *density* of negatives.

Let extracted representations of negative points be $\mathcal{T}_{\text{neg}}$. When combining negative points into contrastive learning, the similarity matrix of Eq. 9 is modified to $\mathbf{J}(\mathcal{T}_s, \mathcal{T}_{s'} \cup \mathcal{T}_{\text{neg}})$.

Optimization

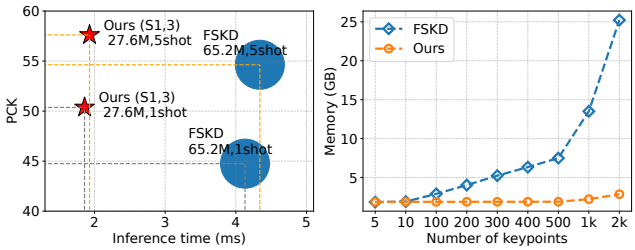
Considering the heatmap regression loss $\mathcal{L}_{\text{hm}}$ and the MFCL loss $\mathcal{L}_{\text{CL}}^{N_e}$, we obtain the overall loss given as

$$\mathcal{L} = \mathcal{L}_{\text{hm}} + \lambda \mathcal{L}_{\text{CL}}^{N_e}, \tag{13}$$

[1] $\mathcal{B}_{1.0}$ is computed without using extra human annotations.

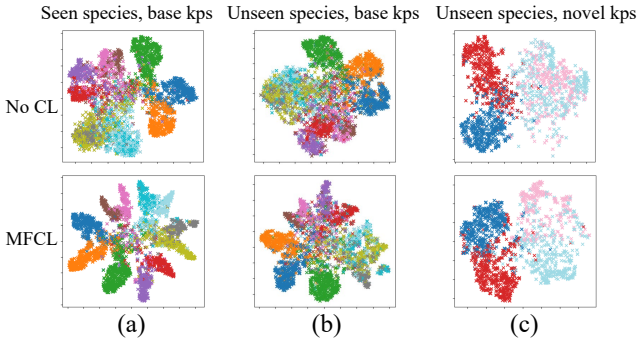
Model	Params	IT	Train Time	Memory	Mean PCK
FSKD	65.2M	4.13 ms	24.1h	1903M	44.75
Ours (S=1)	27.5M (57%↓)	1.76 ms (57%↓)	6.7h (72%↓)	1435M (24%↓)	49.22 (4.47↑)
Ours (S=3)	27.6M (57%↓)	1.81 ms (56%↓)	7.3h (69%↓)	1435M (24%↓)	48.54 (3.79↑)
Ours (S=5)	27.6M (57%↓)	3.36 ms (18%↓)	8.0h (66%↓)	1435M (24%↓)	48.82 (4.07↑)
Ours (S=1,3)	**27.6M (57%↓)**	**1.86 ms (54%↓)**	**7.8h (67%↓)**	**1437M (24%↓)**	**50.37 (5.62↑)**

Table 1: Efficiency test of 1-shot novel keypoint detection on the Animal pose dataset. The mean PCK over 5-subproblems is reported. "IT" is average keypoint inference time per query image measured in V100. "S" means size of generated conv. kernel.



(a) (b)

Figure 4: GPU memory over number of tested keypoints simulated on the 1-shot novel keypoint detection.



(a) (b) (c)

Figure 5: Visualization of the distribution of keypoint representations in different domains. "no CL": don't use contrastive learning; "MFCL": use MFCL loss.

where λ is the loss weight. By default, we set $\lambda = 0.002$. Note that MFCL loss serves as a regularization for the main task loss. Moreover, N_e is usually small and we set N_e to 4. Thus, the optimization of overall loss is efficient.

Experiments

Datasets and Protocols

We evaluate FSKD models using four datasets as follows:

- **Animal pose dataset** (Cao et al. 2019a) has five mammal species, *i.e.*, *cat*, *dog*, *cow*, *horse*, and *sheep*, with over 6000 instances with keypoint annotations. For Animal pose dataset, one animal species is alternately chosen as unseen species for testing while the remaining four as seen species for training, which yields five subproblems.
- **AwA** (Banik, Li, and Dong 2021) has 35 diverse animal species with 10064 images. For AwA, 25 species are for training and 10 for testing.
- **CUB** (Wah et al. 2011) consists of 200 species with 15 keypoint annotations. We use 100 species for training, 50 for validation, and 50 for testing.
- **NABird** (Van Horn et al. 2015) is a larger dataset than CUB with 555 categories, 11 types of annotated body parts, and 48,562 images. The species split is 333, 111, and 111 for training, validation and testing respectively.

Moreover, we also follow Lu and Koniusz (2022) and split keypoints into base and novel sets, and report the performance on both novel and base keypoints.

Experimental Setup

Metric: We use the percentage of correct keypoints. A predicted keypoint is correct if its distance to GT $d \leq \tau \cdot \max(w_{\text{bbx}}, h_{\text{bbx}})$, where w_{bbx} and h_{bbx} are the edges of object bounding box, and $\tau = 0.1$.

Implementation Details: The input image size for all models is 384×384 and the backbone of all compared methods uses ResNet50 (He et al. 2016). By default, our non-linear KG uses two resolutions $S = \{1, 3\}$, *i.e.*, the group number of kernels and heatmaps is 2. In MFCL, the temperature $\tau = 0.05$. The setting of negative keypoints is $(\alpha, \rho, N_{\text{neg}}) = (1.15, 30, 10)$.

Compared Methods: As previous works, we use for comparisons few-shot learning models *ProtoNet* (Snell, Swersky, and Zemel 2017), *RelationNet* (Sung et al. 2018), and *LwoF* (with or w/o attention) (Gidaris and Komodakis 2018), and FSKD-dedicated works *FSKD-R/-D* (Lu and Koniusz 2022). Our method is denoted as *lwFSKD*.

Efficiency Test Results

Below, we configure our lightweight model with different resolutions of generated conv. kernels in non-linear KG. Table 1 shows that using both 1×1 and 3×3 kernels is best, as suitable multi-group kernels improve the correlation window during SMD and alleviate the scale issue of query objects. Compared to FSKD-D, our model enjoys over 50% parameters reduction and achieves 5.62% improvements on the 1-shot novel keypoint detection on the Animal pose dataset.

Fig. 4a shows our model strongly surpasses FSKD model proposed by (Lu and Koniusz 2022) in both 1-shot and 5-shot setting on animal dataset. Moreover, Fig. 4b highlights that our method significantly reduces memory usage, and sheds light on testing on given large numbers of keypoints.

Setting	Model	Params	IT	Animal Pose Dataset						AwA	CUB	NABird
				Cat	Dog	Cow	Horse	Sheep	Avg			
Novel	ProtoNet	19.0M	3.65 ms	19.68	16.18	14.39	12.05	15.06	15.47	29.57	51.32	36.65
	RelationNet	21.1M	3.74 ms	22.15	17.19	15.47	13.58	16.55	16.99	20.91	56.59	34.02
	LwoF (w/o Att.)	19.0M	3.81 ms	21.86	17.11	16.19	16.34	16.13	17.53	28.13	52.66	33.31
	LwoF	23.2M	4.01 ms	22.47	19.39	16.82	16.40	16.94	18.40	28.54	54.75	34.19
	ProbIntr	29.5M	3.84 ms	28.54	23.20	19.55	17.94	17.03	21.25	32.00	68.07	48.70
	FSKD-R	65.2M	4.13 ms	46.05	40.66	37.55	38.09	31.50	38.77	51.81	77.90	54.01
	FSKD-D	65.2M	4.13 ms	52.36	47.94	44.07	42.77	36.60	44.75	64.76	77.89	56.04
	lwFSKD (S=1)	27.5M	1.76 ms	57.53	51.55	49.39	45.26	42.38	49.22	67.46	80.63	57.87
	lwFSKD (S=3)	27.6M	1.81 ms	55.86	51.25	48.28	46.09	41.23	48.54	68.05	82.10	58.18
	lwFSKD (S=5)	27.6M	3.36 ms	55.58	50.60	49.00	45.81	41.75	48.82	67.14	80.42	58.65
	lwFSKD (S=1,3)	**27.6M**	**1.86 ms**	**59.03**	**52.03**	**49.79**	**47.70**	**43.30**	**50.37**	**69.81**	**83.34**	**60.92**
Base	ProtoNet	19.0M	3.56 ms	45.80	39.83	34.88	35.80	32.33	37.73	57.17	80.36	73.18
	RelationNet	21.1M	3.33 ms	51.03	45.85	39.86	41.97	37.19	43.18	57.31	79.40	78.85
	LwoF (w/o Att.)	19.0M	3.99 ms	51.52	45.50	43.38	40.15	37.89	43.69	62.77	80.42	80.83
	LwoF	23.2M	3.79 ms	50.05	44.64	43.47	43.35	37.84	43.87	63.87	81.96	81.39
	ProbIntr	29.5M	3.70 ms	45.96	42.49	37.87	40.53	37.04	40.78	58.51	73.46	70.56
	FSKD-R	65.2M	3.67 ms	57.12	51.12	47.83	49.71	43.71	49.90	65.26	87.94	87.84
	FSKD-D	65.2M	3.67 ms	56.38	51.29	48.24	49.77	43.95	49.93	66.39	87.71	86.99
	lwFSKD (S=1)	27.5M	1.53 ms	55.93	53.87	52.69	54.39	47.25	52.83	68.74	92.52	90.33
	lwFSKD (S=3)	27.6M	1.62 ms	57.43	53.99	53.88	53.04	48.53	53.37	71.73	90.80	89.42
	lwFSKD (S=5)	27.6M	3.17 ms	58.45	53.21	52.78	53.43	47.28	53.03	70.14	93.20	89.97
	lwFSKD (S=1,3)	**27.6M**	**1.63 ms**	**58.64**	**55.44**	**54.05**	**55.28**	**49.87**	**54.66**	**71.86**	**94.16**	**91.81**

Table 2: Results on the 1-shot keypoint detection for unseen species across four datasets. The PCK scores on novel and base keypoint detection are reported. "IT" is average keypoint inference time measured on the Animal pose dataset via a V100 GPU.

Setting	Model	Params	Animal	AwA	CUB	NABird
Novel	ProtoNet	19.0M	23.66	36.94	61.17	52.56
	RelationNet	21.1M	21.37	28.58	60.41	41.42
	LwoF (w/o Att.)	19.0M	21.92	35.23	61.77	47.78
	LwoF	23.2M	27.70	40.02	60.70	48.03
	ProbIntr	29.5M	34.31	51.68	75.65	60.86
	FSKD-R	65.2M	51.42	68.39	80.26	63.76
	FSKD-D	65.2M	54.63	74.47	79.17	63.46
	lwFSKD	**27.6M**	**57.61**	**77.23**	**80.74**	**67.79**

Table 3: Results on the 5-shot novel keypoint detection for unseen species across four datasets. We report PCK scores.

One-shot Novel Keypoint Det.	Animal	AwA	CUB	NABird
0 *Baseline* ($\diamond$)	39.82	59.96	74.03	49.16
1 $\diamond$+$\mathcal{U}$	41.71	62.98	77.02	54.11
2 $\diamond$+$\mathcal{U}$+*linear KG*	43.01	63.13	77.61	54.69
4 $\diamond$+$\mathcal{U}$+*non-linear KG*	46.04	66.65	79.36	57.54
5 $\diamond$+$\mathcal{U}$+*non-linear KG*+*CL*	47.63	67.72	80.37	57.83
6 $\diamond$+$\mathcal{U}$+*non-linear KG*+*MFCL*†	49.39	68.43	82.48	59.91
7 $\diamond$+$\mathcal{U}$+*non-linear KG*+*MFCL*	**50.37**	**69.81**	**83.34**	**60.92**

Table 4: Ablation study. *CL* denotes instance-based contrastive learning; *MFCL*† means w/o negative keypoints.

Results on Few-shot Keypoint Detection

Below, we conduct comprehensive experiments on few-shot keypoint detection across four datasets.

Firstly, we perform 1-shot keypoint detection for unseen species across the Animal pose dataset, AwA, CUB, and NABird. Table 2 shows our lightweight model consistently achieves best scores, which highlights the efficacy of proposed few-shot model. Secondly, we increase the number of shots, and evaluate the 5-shot novel keypoint detection. Table 3 shows the scores improve further, *e.g.*, compare 57.61% *vs*. 50.37% in the Animal pose dataset. Again, our model outperforms other methods.

Analysis. Table 4 shows the ablations with or w/o upsampling module $\mathcal{U}$, no KG/linear KG/non-linear KG, instance-based CL, and mean feature based contrastive learning (MFCL) with and w/o negative keypoints. We discover that each module is effective in our FSKD system.

Visualization. We also visualize the distribution of keypoint representations via t-SNE (Van der Maaten and Hinton 2008). As shown in Fig. 5, the boundary of different body parts becomes clearer after applying MFCL.

Conclusion

We propose an extremely lightweight few-shot keypoint detector. We theoretically analyze that the design of simultaneous modulation and detection with our non-linear kernel generator and non-parametric detection can reduce cost while achieves compelling results supported by experiments. Moreover, our MFCL further improves few-shot model. We believe the lightweight FSKD and MFCL will provide useful insights on general keypoint detection, thus we highly recommend it to vision and learning community.

References

Banik, P.; Li, L.; and Dong, X. 2021. A Novel Dataset for Keypoint Detection of quadruped Animals from Images. *arXiv preprint arXiv:2108.13958*.

Bardes, A.; Ponce, J.; and LeCun, Y. 2021. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*.

Bohdal, O.; Tian, Y.; Zong, Y.; Chavhan, R.; Li, D.; Gouk, H.; Guo, L.; and Hospedales, T. 2023. Meta Omnium: A Benchmark for General-Purpose Learning-to-Learn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7693–7703.

Cao, J.; Tang, H.; Fang, H.-S.; Shen, X.; Lu, C.; and Tai, Y.-W. 2019a. Cross-domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9498–9507.

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2019b. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1): 172–186.

Carreira, J.; Agrawal, P.; Fragkiadaki, K.; and Malik, J. 2016. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4733–4742.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T. S.; and Zhang, L. 2020. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5386–5395.

Derpanis, K. G. 2004. The harris corner detector. *York University*, 2–3.

Doersch, C.; Gupta, A.; and Zisserman, A. 2020. Crosstransformers: spatially-aware few-shot transfer. *Advances in Neural Information Processing Systems*, 33: 21981–21993.

Fan, Q.; Zhuo, W.; Tang, C.-K.; and Tai, Y.-W. 2020. Few-shot object detection with attention-RPN and multi-relation detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4013–4022.

Fang, H.-S.; Xie, S.; Tai, Y.-W.; and Lu, C. 2017. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, 2334–2343.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 1126–1135. PMLR.

Ge, Y.; Zhang, R.; and Luo, P. 2021. MetaCloth: Learning Unseen Tasks of Dense Fashion Landmark Detection From a Few Samples. *IEEE Transactions on Image Processing*, 31: 1120–1133.

Gidaris, S.; and Komodakis, N. 2018. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4367–4375.

Graving, J. M.; Chae, D.; Naik, H.; Li, L.; Koger, B.; Costelloe, B. R.; and Couzin, I. D. 2019. DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife*, 8: e47994.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, X.; Bharaj, G.; Ferman, D.; Rhodin, H.; and Garrido, P. 2023. Few-shot Geometry-Aware Keypoint Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21337–21348.

Hou, R.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2019. Cross attention network for few-shot classification. *Advances in neural information processing systems*, 32.

Jian, Y.; Gao, C.; and Vosoughi, S. 2022. Contrastive learning for prompt-based few-shot language learners. *arXiv preprint arXiv:2205.01308*.

Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; and Darrell, T. 2019. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8420–8429.

Koch, G.; Zemel, R.; Salakhutdinov, R.; et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille.

Kowalski, M.; Naruniec, J.; and Trzcinski, T. 2017. Deep alignment network: A convolutional neural network for robust face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 88–97.

Liu, C.; Fu, Y.; Xu, C.; Yang, S.; Li, J.; Wang, C.; and Zhang, L. 2021. Learning a few-shot embedding model with contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 8635–8643.

Lowe, G. 2004. SIFT-the scale invariant feature transform. *Int. J*, 2: 91–110.

Lu, C.; and Koniusz, P. 2022. Few-shot keypoint detection with uncertainty learning for unseen species. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19416–19426.

Lu, C.; Zhu, H.; and Koniusz, P. 2023. From Saliency to DINO: Saliency-guided Vision Transformer for Few-shot Keypoint Detection. *arXiv preprint arXiv:2304.03140*.

Moskvyak, O.; Maire, F.; Dayoub, F.; and Baktashmotlagh, M. 2021. Semi-supervised keypoint localization. *arXiv preprint arXiv:2101.07988*.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, 483–499. Springer.

Ouali, Y.; Hudelot, C.; and Tami, M. 2021. Spatial contrastive learning for few-shot classification. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part I 21*, 671–686. Springer.

Pereira, T. D.; Aldarondo, D. E.; Willmore, L.; Kislin, M.; Wang, S. S.-H.; Murthy, M.; and Shaevitz, J. W. 2019. Fast animal pose estimation using deep neural networks. *Nature methods*, 16(1): 117–125.

Shi, M.; Huang, Z.; Ma, X.; Hu, X.; and Cao, Z. 2023a. Matching Is Not Enough: A Two-Stage Framework for Category-Agnostic Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7308–7317.

Shi, W.; Lu, C.; Shao, M.; Zhang, Y.; Xia, S.; and Koniusz, P. 2023b. Few-shot Shape Recognition by Learning Deep Shape-aware Features. *arXiv preprint arXiv:2312.01315*.

Snell, J.; Swersky, K.; and Zemel, R. S. 2017. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*.

Sohn, K. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5693–5703.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1199–1208.

Tang, L.; Wertheimer, D.; and Hariharan, B. 2020. Revisiting pose-normalization for fine-grained few-shot recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14352–14361.

Tompson, J. J.; Jain, A.; LeCun, Y.; and Bregler, C. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems*, 27.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Van Horn, G.; Branson, S.; Farrell, R.; Haber, S.; Barry, J.; Ipeirotis, P.; Perona, P.; and Belongie, S. 2015. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 595–604.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29: 3630–3638.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Weinberger, K. Q.; and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2).

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

Xu, L.; Jin, S.; Zeng, W.; Liu, W.; Qian, C.; Ouyang, W.; Luo, P.; and Wang, X. 2022. Pose for everything: Towards category-agnostic pose estimation. In *European Conference on Computer Vision*, 398–416. Springer.

Yang, Z.; Wang, J.; and Zhu, Y. 2022. Few-shot classification with contrastive learning. In *European Conference on Computer Vision*, 293–309. Springer.

Yao, Q.; Quan, Q.; Xiao, L.; and Kevin Zhou, S. 2021. One-shot medical landmark detection. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, 177–188. Springer.

Yosida, K. 2012. *Functional analysis*. Springer Science & Business Media.

Zhang, S.; Luo, D.; Wang, L.; and Koniusz, P. 2020. Few-shot object detection by second-order pooling. In *ACCV*.

Zhao, S.; Gong, M.; Zhao, H.; Zhang, J.; and Tao, D. 2023. Deep Corner. *International Journal of Computer Vision*, 1–25.