Few-shot Keypoint Detection with Uncertainty Learning for Unseen Species (Supplementary Material)

Changsheng Lu[†], Piotr Koniusz[*,§,†]
[†]The Australian National University [§]Data61/CSIRO
ChangshengLuu@gmail.com, firstname.lastname@anu.edu.au

Our code will be released in the future: https://github.com/AlanLuSun/Few-shot-keypoint-detection.

A. Uncertainty-weighted TPS Warp

A.1. Proof of the Solution

Let us denote two sets of fiducial points as $\mathbf{P} = [\mathbf{p}_1, \cdots, \mathbf{p}_N] \in \mathbb{R}^{2\times N}$ and $\mathbf{P}' = [\mathbf{p}'_1, \cdots, \mathbf{p}'_N] \in \mathbb{R}^{2\times N}$, where $\Pi_i = (\mathbf{p}_i, \mathbf{p}'_i)$ is a pair of corresponding points. Let $\mathbf{W} = diag([w_1, \cdots, w_N]) \in \mathbb{S}^N_{++}$ be a diagonal matrix whose each diagonal entry w_i is the confidence for Π_i. Then, our goal is to find a function mapping $f : \mathbf{p}_i \rightarrow \mathbf{p}'_i$ to achieve the minimal distance error between $\mathbf{P}$ and $\mathbf{P}'$ while ensuring the least deformation in rigidity. The objective of weighted TPS warp can be formulated as

$$E(f, \mathbf{P}, \mathbf{P}') = \sum_{i=1}^{N} w_i^2 \|\mathbf{p}'_i - f(\mathbf{p}_i)\|_2^2 + \lambda \iint_{[x,y]^\intercal \in \mathbb{R}^2} \left(\frac{\partial^2 f}{\partial x^2}\right)^2 + \left(\frac{\partial^2 f}{\partial y^2}\right)^2 + 2\left(\frac{\partial^2 f}{\partial x \partial y}\right)^2 dxdy, \tag{12}$$

where the former term describes the weighted distance error, and the latter term (the definite integral) penalizes the so-called bending energy. Function f can be constructed using the combination of affine transformation and a set of radial basis functions (RBF) as

$$f(\mathbf{p}) = \mathbf{a}_1 + \mathbf{a}_2 x + \mathbf{a}_3 y + \sum_{i=1}^{N} \mathbf{b}_i \phi(\|\mathbf{p} - \mathbf{p}_i\|_2), \tag{13}$$

where $\mathbf{a}_i, \mathbf{b}_i \in \mathbb{R}^2$, $\mathbf{p} = [x, y]^\intercal \in \mathbb{R}^2$, and $\phi(\cdot)$ is a function with the radial basis. Following [3], when choosing $\phi(d) = d^2 \log d^2$, the bending energy in Eq. 12 can be minimized and the objective function could be reduced as

$$E(f, \mathbf{P}, \mathbf{P}') = \sum_{i=1}^{N} w_i^2 \|\mathbf{p}'_i - \mathbf{A}\hat{\mathbf{p}}_i - \mathbf{B}\boldsymbol{\gamma}_i\|_2^2 + \lambda \text{tr}(\mathbf{B}\mathbf{R}\mathbf{B}^\intercal) \tag{14}$$

[*]The corresponding author.

where $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3] \in \mathbb{R}^{2\times 3}$, $\mathbf{B} = [\mathbf{b}_1, \cdots, \mathbf{b}_N] \in \mathbb{R}^{2\times N}$. $\hat{\mathbf{p}}_i = [1, \mathbf{p}_i^\intercal]^\intercal \in \mathbb{R}^3$ is the homogeneous coordinate. $\boldsymbol{\gamma}_i = [\gamma_{1,i}, \cdots, \gamma_{N,i}]^\intercal \in \mathbb{R}^N$ is a column vector containing the entry terms $\gamma_{n,i} = d_{n,i}^2 \log d_{n,i}^2$, and $d_{n,i}$ is the Euclidean distance between $\mathbf{p}_n$ and $\mathbf{p}_i$. $\mathbf{R} = [\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_N] \in \mathbb{S}^N_{++}$ is a symmetric positive definite matrix, *i.e.*, $\mathbf{R} = \mathbf{R}^\intercal$ and $\mathbf{R} \succ 0$. By differentiating $E(f, \mathbf{P}, \mathbf{P}')$ w.r.t. $\mathbf{A}$ and $\mathbf{B}$, we have

$$\begin{aligned} \frac{\partial E}{\partial \mathbf{A}} &= \sum_{i=1}^{N} w_i^2 [-2\mathbf{p}'_i \hat{\mathbf{p}}_i^\intercal + 2\mathbf{B}\boldsymbol{\gamma}_i \hat{\mathbf{p}}_i^\intercal + 2\mathbf{A}\hat{\mathbf{p}}_i \hat{\mathbf{p}}_i^\intercal] \\ \frac{\partial E}{\partial \mathbf{B}} &= \sum_{i=1}^{N} w_i^2 [-2\mathbf{p}'_i \boldsymbol{\gamma}_i^\intercal + 2\mathbf{A}\hat{\mathbf{p}}_i \boldsymbol{\gamma}_i^\intercal + 2\mathbf{B}\boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^\intercal] + 2\lambda \mathbf{B}\mathbf{R} \end{aligned} \tag{15}$$

Let $\frac{\partial E}{\partial \mathbf{A}} = \mathbf{0}$ and $\frac{\partial E}{\partial \mathbf{B}} = \mathbf{0}$, we obtain the constraints as follows

$$(-\mathbf{P}' + \mathbf{B}\mathbf{R} + \mathbf{A}\hat{\mathbf{P}})\mathbf{W}^2 \hat{\mathbf{P}}^\intercal = \mathbf{0} \tag{16}$$

$$\mathbf{A}\hat{\mathbf{P}} + \mathbf{B}(\mathbf{R} + \lambda \mathbf{W}^{-2}) = \mathbf{P}', \tag{17}$$

where $\hat{\mathbf{P}} = [\hat{\mathbf{p}}_1, \cdots, \hat{\mathbf{p}}_N] \in \mathbb{R}^{3\times N}$. By substituting $\mathbf{P}'$ in Eq. 16 with Eq. 17, we have

$$\mathbf{B}\hat{\mathbf{P}}^\intercal = \mathbf{0}. \tag{18}$$

Using Eq. 17 and Eq. 18, we build the linear system as

$$\begin{bmatrix} \mathbf{R} + \lambda \mathbf{W}^{-2} & \hat{\mathbf{P}}^\intercal \\ \hat{\mathbf{P}} & \mathbf{0}^{3\times 3} \end{bmatrix} \begin{bmatrix} \mathbf{B}^\intercal \\ \mathbf{A}^\intercal \end{bmatrix} = \begin{bmatrix} \mathbf{P}'^\intercal \\ \mathbf{0}^{3\times 2} \end{bmatrix}. \tag{19}$$

Solving the Eq. 19, the transformation parameters $\mathbf{T} = [\mathbf{B}, \mathbf{A}] \in \mathbb{R}^{2\times(N+3)}$ can be obtained.

If we have the uncertainty for each pair of points Π_i as J_i, and denote $\mathbf{D} = diag([J_1, \cdots, J_N]) \in \mathbb{S}^N_{++}$ and $\mathbf{W} = \mathbf{D}^{-1}$, then the uncertainty-weighted TPS warp can be solved using Eq. 19. The proof is completed.

A.2. Toy Experiment

To validate the effectiveness of uncertainty-weighted TPS warp, we warp the image given the corresponding keypoints as shown in Fig. 10(a), where the tilted crosses are

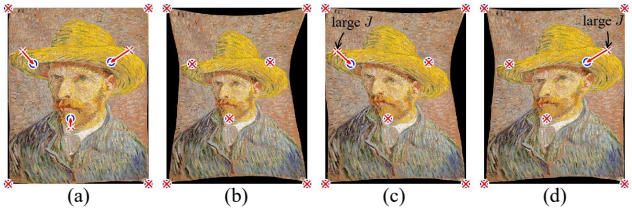

Figure 10. An example of using uncertainty-weighted TPS warp. (a) Original image that contains two sets of corresponding keypoints marked by red tilted crosses and blue circles; (b) perfect warping; (c) large uncertainty J for the left keypoint on the hat; (d) large uncertainty J for the right keypoint on the hat.

source keypoints and blue circles are target keypoints. After warping, the source keypoints will move to target keypoints. We manually set one keypoint correspondence to be with large uncertainty strength $J = 100$ while other keypoints are with low uncertainty $J = 1$. As shown in Fig. 10(c) and Fig. 10(d), the keypoint with the large uncertainty J is less warped than other keypoints, which indicates that the proposed approach enjoys larger tolerance to the uncertain keypoints by focusing more on those confident keypoints.

B. Further Details for Experiment Setup

B.1. Compared Methods

When modifying *ProbIntr* [36] to adapt it to the few-shot keypoint detection task, we also build support keypoint prototypes (SKP) from extracted keypoint representations. The semantic distinctiveness (SD) is learnt with the goal of constructing the probabilistic introspection matching loss $\mathcal{L}_{\text{m}}$ between SKP and individual query feature vectors. In addition to the matching loss from positive pair of keypoints, we also adopt negative pair of keypoints to perform hard negative mining. Moreover, we augment two views for each image in the episode and add the self-supervised loss $\mathcal{L}_{\text{ssl}}$ into the training step by randomly sampling 20 keypoints. The adopted feature encoder of *ProbIntr* is ResNet50 which is identical to the encoder employed by our FSKD models. The whole network is trained by jointly optimizing $\mathcal{L} = \alpha\mathcal{L}_{\text{m}} + \mathcal{L}_{\text{ssl}}$, where α is set experimentally to 0.075 (for the best performance of that baseline).

In our FSKD architecture, the output feature map of encoder has the size of $2048 \times 12 \times 12$ which indicates the downsize factor of $f = \frac{1}{32}$ compared to the image length. Since the model pretrained on ImageNet [10] provides stable low-level features and helps convergence, we fix the weights of the first three convolutional (conv.) blocks of encoder. When using Gaussian pooling to extract keypoint representations, we set $\xi = 14f = \frac{14}{32}$. The SD head consists of two conv. layers and a 1×1 conv. filter to convert the intermediate features into a single-channel SD map $\boldsymbol{\sigma}^{-1}$. We perform numerical transformation $f(x) = \frac{1}{2}(x + \sqrt{x^2 + \epsilon})$ to ensure SD map $\boldsymbol{\sigma}^{-1} > 0$. The input and output of descriptor extractor contain dedicated conv. layers in order to manipulate their feature maps to desired sizes, whereas the intermediate layers contain a series of 3×3 conv. blocks which continuously reduce the feature map size. In our UC-GBL, all branches are implemented with MLP. We use the Adam optimizer and set the learning rate to $1e-4$.

Table 5. Keypoint splits used in our experiments for three datasets.

Dataset	Base Keypoint Set	Novel Keypoint Set
Animal	*two ears, nose, four legs, four paws*	*two eyes, four knees*
CUB	*beak, belly, back, breast, crown, two legs, nape, throat, tail*	*forehead, two eyes, two wings*
NABird	*beak, belly, back, breast, crown, nape, tail*	*two eyes, two wings*

Table 6. Additional comparison results on 1-shot novel keypoint detection.

Method	Animal Pose Dataset						CUB	NABird
	Cat	Dog	Cow	Horse	Sheep	Avg		
Baseline	27.30	24.40	19.40	18.25	21.22	22.11	66.12	39.14
ProbIntr [36]	28.54	23.20	19.55	17.94	17.03	21.25	68.07	48.70
TFA [60]	19.40	20.00	20.85	17.99	19.54	19.56	50.12	30.16
ProtoNet [42]	19.68	16.18	14.39	12.05	15.06	15.47	51.32	36.65
RelationNet [45]	22.15	17.19	15.47	13.58	16.55	16.99	56.59	34.02
WG (w/o Att.) [59]	21.86	17.11	16.19	16.34	16.13	17.53	52.66	33.31
WG [59]	22.47	19.39	16.82	16.40	16.94	18.40	54.75	34.19
FSKD (rand) (Ours)	46.05	40.66	37.55	38.09	31.50	38.77	77.90	54.01
FSKD (default) (Ours)	52.36	47.94	44.07	42.77	36.60	44.75	77.89	56.04

B.2. Detailed Keypoint Splits

We split the base keypoint set and the novel keypoint set of each dataset as detailed in Table 5. These splits are used in our experiments. We notice that other split choices could also be used in our FSKD pipeline.

C. Additional FSKD Results

Popular few-shot learning (FSL) methods *e.g.*, *ProtoNet* [42], *RelationNet* [45], two versions of *WG* (with or without attention) [59], and Two-stage Finetuning Approach (*TFA*) [60] are adapted to perform the FSKD task. All methods use the ResNet50 backbone and are evaluated under the setting of 1-shot novel keypoint detection (**Sec.** 4.2). Table 6 shows that the adapted state-of-the-art FSL approaches struggle to learn from the limited number of base keypoints. In constrast, thanks to the novel FSKD-specific designs such as single/multi-keypoint uncertainty modeling, auxiliary keypoints learning, and multi-scale UC-GBL, our FSKD variants achieve the best performance and outperform the above baselines by a large margin.

Figure 11. Extensive examples of 1-shot detection for novel keypoints in unseen species. From (a)∼(e), each row is a subproblem by regarding an animal as unseen species in animal pose dataset, which is cat, dog, cow, horse, and sheep; (f) and (g) are results from 1-shot tasks in CUB and NABird, respectively. The experiments run in same-species episodes. The novel keypoint predictions (tilted crosses), estimated localization uncertainty (red ellipses), and groundtruth keypoints (circles) are simultaneously drawn.

Moreover, we visualize additional 1-shot detection results for novel keypoints in unseen species. As shown in Fig. 11, despite the query images containing various detrimental factors such as numerous behaviors, complex natural backgrounds, shadows, and areas of low contrast, the proposed FSKD successfully detects novel keypoints in each query image given the support keypoints. Further, the estimated uncertainty marked by the red ellipse covers both the keypoint prediction and GT location, which indicates that the localization uncertainty is a good indicator of where the possible GT keypoint is located. Interestingly, the uncertainty distribution exhibits a relationship with the shape of body parts, which should help limit the ambiguity of keypoints.

D. Additional Ablation Study

In this section, we present more ablation studies to validate the effectiveness of components involved in our pipeline. Similar to Section 4.3 in the paper, we use the all-way-1-shot novel keypoints detection in unseen species with FSKD (default) running on same-species episodes.

Additional Results on Multi-scale UC-GBL: We visualize outputs of each scale from Multi-scale UC-GBL, with an example shown in Fig. 12. Indeed, the keypoint prediction from multi-scale UC-GBL is more stable, reducing the

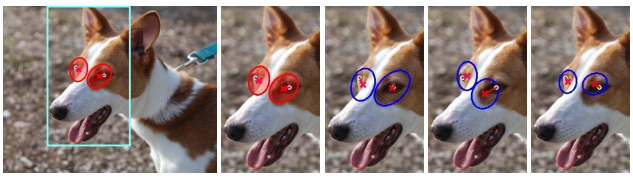

Figure 12. MS UC-GBL decomposition and uncertainty fusion.

Table 7. Study on keypoint feature extraction strategies and improvements by self-modulation during meta-testing.

Extraction method	Self-modulation	Cat	Dog
Integer-based indexing	✗	48.11	40.60
Bilinear interpolation	✗	52.30	47.18
Gaussian pooling	✗	**52.36**	**47.94**
Gaussian pooling	*1 gradient-step*	52.88	48.96
Gaussian pooling	*2 gradient-step*	53.66	49.08
Gaussian pooling	*3 gradient-step*	54.01	**49.40**
Gaussian pooling	*4 gradient-step*	**54.43**	49.00
Gaussian pooling	*5 gradient-step*	52.87	49.36

Figure 13. Visualization of semantic distinctiveness map $\boldsymbol{\sigma}^{-1}$.

Table 8. Results on 1-shot keypoint detection for unseen species. The mix-species episode is used.

Setting	Method	CUB	NABird
Novel	Baseline	65.45	34.48
	ProbIntr	59.07	35.06
	FSKD (rand)	75.27	50.25
	FSKD (default)	**76.99**	**51.22**
Base	Baseline	80.81	74.10
	ProbIntr	71.40	74.82
	FSKD (rand)	86.92	80.25
	FSKD (default)	**87.66**	**84.74**

risk of mislocalization. Increasing scale S makes the grid finer and thus the uncertainty range shrinks. However, our fused uncertainty yields a good quality of combined uncertainty estimation.

Body Part Extraction Strategies: We compare the impacts of three keypoint feature extraction approaches such as the integer-based indexing, bilinear interpolation, and Gaussian pooling, given the same architecture, and perform experiments on two subproblems by regarding cat and dog as unseen species, respectively. Table 7 shows that the bilinear interpolation and Gaussian pooling yield better results as the extracted soft keypoint representations contain larger spatial context, which helps build a more expressive support keypoint prototype (SKP).

Self-modulation in Meta-testing: In addition, we find that the learnt FSKD model can improve its performance via self-modulation during meta-testing. Following Model-Agnostic Meta-Learning (MAML) [16], in each episode, we fine-tune the learnt meta-model via several gradient descent steps of back-propagation such that the meta-model has a chance to adapt better to the test data. Specifically, given the access to the support keypoints in the support image during meta-testing, we use SKPs to modulate the *support feature map* and construct the loss using *support keypoints*. After several gradient descent steps of back-propagation, the fine-tuned model is used for detecting the corresponding keypoints in the query image by modulating the *query feature map*. Table 7 shows significant gains when using self-modulation, *e.g.*, 54.43% (4 gradient-step fine-tuning) *vs*. 52.36% (without fine-tuning) for the cat. Meanwhile, we note that the excessive fine-tuning may overfit.

Semantic Distinctiveness Map: Examples of semantic distinctiveness (SD) map $\boldsymbol{\sigma}^{-1}$ are shown in Fig. 13.

Mix-species Episode: To investigate the few-shot keypoint detection with the mix-species episodes, we perform the experiments on CUB and NABird with the goal of detecting keypoints in unseen species. Table 8 shows that the proposed FSKD variants are still effective but incur slight performance drops compared to the results in keypoint detection using same-species episodes, *e.g.*, 51.22% (Table 8) *vs*. 56.04% (Main paper, Table 1) achieved by FSKD (default) in NABird. The mix-species episode leads to a larger domain shift between the support and query images and thus poses an additional challenge to the model learning and keypoint localization.

E. Visualizations of Semantic Alignment

Extensive qualitative results of semantic alignment (SA) are shown in Fig. 14. We perform SA for unseen species using 1-shot FSKD model trained on mix-species episodes. The reason we chose the mix-species episode setting is that aligning objects of different visual categories yields more diverse SA results in this setting.

When performing *Warp with GT* query keypoints (Fig. 14, 3rd column), even though most query keypoints (marked by tilted crosses) align perfectly with the support keypoints (marked by circles), *Warp with GT* results in unacceptable deformations of objects. In contrast, *Identical UC* (Fig. 14, 4th column) maintains the shape relatively

Figure 14. Additional qualitative results of semantic alignment using different approaches. The first column shows the support keypoints & image; the second column shows the query image with the predicted keypoints (marked by tilted crosses) and uncertainty (red shadow ellipses); the last three columns are the results achieved by *Warp with GT* [3], *Identical UC*, and our uncertainty-weighted TPS warp.

better compared to *Warp with GT* by applying the identical warping penalty. However, as weights of warping penalization are equal across keypoints, one can see that the deformations may appear in the proximity of inaccurate or poorly corresponding keypoints. In contrast to *Warp with GT* and *Identical UC*, our uncertainty-weighted TPS warp addresses the above issues, thus producing a much better perceptual alignment. Additionally, despite the support and query images are from different species and often have very differed poses, our FSKD detects the keypoints reliably, estimates the uncertainty reliably, and thus leads to a high-quality semantic alignment.

F. Discussion

Difference Compared with Other Few-shot Tasks: Compared to few-shot image classification (FSL) and few-shot object detection (FSOD), there are *two* main difference in FSKD.

Firstly, in FSL and FSOD, N*-way learning* refers to N visual categories in support set, while in FSKD, N-way means there are N different keypoint types. Secondly, the *training & testing splits* are devised differently in FSKD, as detailed below.

Denote the set of classes of the training species as $C = \{c_i\}_{i=1,2,\cdots,N_C}$ and the set of classes of testing species as $C' = \{c'_i\}_{i=1,2,\cdots,N_{C'}}$, where each element represents a class label. Let the set of training keypoint types be

$\mathcal{X} = \{k_i\}_{i=1,2,\cdots,N_{\mathcal{X}}}$ and the set of testing keypoint types be $\mathcal{X}' = \{k'_i\}_{i=1,2,\cdots,N_{\mathcal{X}'}}$.

In FSL and FSOD, one splits the species into base and novel class to guarantee $C \cap C' = \varnothing$. While in FSKD, one needs to *split both species and keypoint types* and consider the following possible settings:

- $C \cap C' = \varnothing$ and $\mathcal{X} \cap \mathcal{X}' = \varnothing$, (the hardest setting)
- $C = C'$ and $\mathcal{X} \cap \mathcal{X}' = \varnothing$, (intermediate difficulty)
- $C \cap C' = \varnothing$ and $\mathcal{X} = \mathcal{X}'$, (intermediate difficulty)
- $C = C'$ and $\mathcal{X} = \mathcal{X}'$. (the easiest setting)

Limitations and the Future Work: Learning from several annotated samples to detect novel keypoints is hard but it could be improved with more expressive keypoint representations beyond the Gaussian pooling and advanced feature modulation schemes. Furthermore, the auxiliary keypoints interpolated along lines are suboptimal due to their imprecise matching relationship in locations between support and query, which yields a relatively large localization noise. Despite we address the impact of noise via our uncertainty modeling, we hope to improve the signal-to-noise ratio with more advanced interpolation strategies that could take each shape of object into account.

References

[59] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4367–4375, 2018. 13

[60] Xin Wang, Thomas Huang, Joseph Gonzalez, Trevor Darrell, and Fisher Yu. Frustratingly simple few-shot object detection. In *International Conference on Machine Learning*, pages 9919–9928. PMLR, 2020. 13