Data-Scarce Animal Face Alignment via Bi-Directional Cross-Species Knowledge Transfer

Dan Zeng[*]
Southern University of Science and Technology
Shenzhen, Guangdong, China

Shanchuan Hong[*]
Southern University of Science and Technology
Shenzhen, Guangdong, China

Shuiwang Li
Guilin University of Technology
Guilin, Guangxi, China

Qiaomu Shen[†*]
Southern University of Science and Technology
Shenzhen, Guangdong, China

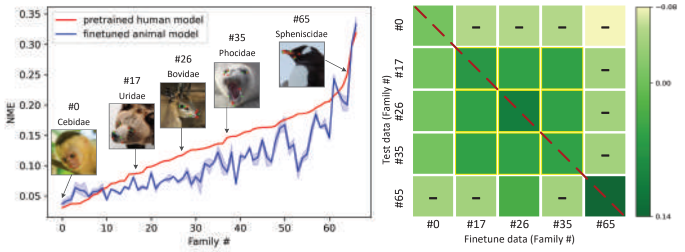
Bo Tang[*]
Southern University of Science and Technology
Shenzhen, Guangdong, China

ABSTRACT

Animal face alignment is challenging due to large intra- and inter-species variations and a scarcity of labeled data. Existing studies circumvent this problem by directly finetuning a human face alignment model or focusing on animal-specific face alignment (e.g., horse, sheep). In this paper, we propose Cross-Species Knowledge Transfer, Meta-CSKT, for animal face alignment, which consists of a base network and an adaptation network. Two networks continuously complement each other through the bi-directional cross-species knowledge transfer. This is motivated by observing knowledge sharing among animals. Meta-CSKT uses a circuit feedback mechanism to improve the base network with the cognitive differences of the adaptation network between few-shot labeled and large-scale unlabeled data. In addition, we propose a positive example mining method to identify positives, semi-hard positives, and hard negatives in unlabeled data to mitigate the scarcity of labeled data and facilitate Meta-CSKT learning. Experiments show that Meta-CSKT outperforms state-of-the-art methods by a large margin on the horse facial keypoint dataset and Japanese Macaque Species dataset, while achieving comparable results to state-of-the-art methods on large-scale labeled AnimalWeb (e.g., 18K), using only a few labeled images (e.g., 40) [1].

CCS CONCEPTS

• **Computing methodologies → Computer vision**; **Biometrics**; **Interest point and salient region detections**; **Semi-supervised learning settings**.



(a) Family-wise Face Alignment Results

(b) NME Difference Confusion Matrix

Figure 1: (a) Results for AnimalWeb by 1 human and 3 animal models, sorted in ascending order of NME by the human model [31]. We randomly select 40 images from AnimalWeb for finetuning and repeat it 3 times. (b) Confusion matrix of NME difference between human and animal models. The x-axis represents the finetune data used to train the animal model, and y-axis represents the test data used to evaluate the model. We randomly select 40 images from *each* family as the finetune data. Elements with/without a minus sign indicate decreased/increased accuracy. Darker Green indicates a larger difference. [Best viewed in color]

KEYWORDS

face alignment; animal biometrics; data scarce; meta learning; semi-supervised learning

ACM Reference Format:
Dan Zeng, Shanchuan Hong, Shuiwang Li, Qiaomu Shen, and Bo Tang. 2023. Data-Scarce Animal Face Alignment via Bi-Directional Cross-Species Knowledge Transfer. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3581783.3612558

1 INTRODUCTION

Animal face alignment aims to detect the facial landmarks on animal faces and has many applications in facial expression analysis, animal pain detection, and facial tracking [1, 21, 22]. The study of animal face alignment is significant as it can help us better understand animals and promote their health by interpreting their facial behavior through visual imagery. This is a less expensive and quicker alternative to clinical examinations and vital signs monitoring. Nonetheless, despite the well-established techniques

[*] Research Institute of Trustworthy Autonomous Systems & Department of Computer Science and Engineering, Southern University of Science and Technology.
[†] **Corresponding author**, Email: shenqm@sustech.edu.cn.
[1] Code required to reproduce our experiment results is available at the link: https://github.com/danzeng1990/Meta-CSKT.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612558

in human face alignment, animal face alignment remains largely unexplored due to large intra- and inter-species variations, as well as the scarcity of labeled data.

The facial appearance of five animals illustrated in Figure 1(a) is subject to significant variability due to both external factors (e.g., illumination, head pose) and internal factors (e.g., animal species, facial expression). The red curve in the figure is biased towards certain species, such as Cebidae (i.e., family 0), that have a close shape or appearance to humans. This bias results in poor accuracy for less human-like species, such as Bovidae (i.e., family 26). Although fine-tuning the human model with a small number of labeled animal images slightly improves the results, the accuracy improvement is uneven across animal families, as demonstrated by the blue curve. Specifically, species at the tail of the curve, such as Spheniscidae (i.e., family 65), show no improvement in accuracy, while species at the head of the curve display an accuracy drop.

Existing studies circumvent this problem by massively increasing the number of labeled animal images [13, 30]. However, obtaining large-scale annotated animal faces can be costly, and training human face alignment networks with such data may be suboptimal as they do not account for inter-species variations (i.e., for human faces, only intra-variations are considered). Some methods focus on animal-specific face alignment [23, 30], such as dogs, sheep, and horses. For example, WarpingNet [23] distorts a horse into a more human-like shape so that the human face alignment model can easily adapt to the horse's appearance. However, this method can only handle species that share certain similarities with humans. In this paper, we study data-scarce animal face alignment where both intra- and inter-species variations are significant, with a focus on using a limited number of labeled animal images.

The confusion matrix of NME difference between the human and animal models is presented in Figure 1(b). We observe that the diagonal elements show increased accuracy. Spheniscidae (i.e., family 65) has the darkest green, and Cebidae (i.e., family 0) has the lightest green on the diagonal elements. This trend is expected, as Spheniscidae is the least human-like species and Cebidae is more similar to humans, resulting in better alignment with the pretrained model. Surprisingly, we also note a similar increased accuracy in the central 3×3 submatrix, clustered in dark green. This finding suggests that these animal species are positively related to each other and share knowledge. These observations motivate Meta-CSKT, a **Meta** optimization framework that leverages **C**ross-**S**pecies **K**nowledge **T**ransfer for animal face alignment. Our animal face alignment is a regression problem, distinct from image classification where labeled and unlabeled images lie in the same feature space and often share the same classification category, and the observed knowledge sharing provides a fundamental premise for Meta-CSKT to function effectively. The proposed method consists of two networks: a base network and an adaptation network, which continuously complement each other in a *bi-directional* manner via CSKT.

In one direction, the adaptation network is trained with large-scale unlabeled data, as knowledge sharing among animals enables the generation of reasonable pseudo ground truths for the unlabeled data. However, due to the gap between a few labeled and large-scale unlabeled data, learning a good adaptation network can be challenging. To this end, we propose feedback learning in the other direction to enhance the base network and refine pseudo labels to improve the adaptation network's performance. The intuition behind the base network update is the relationship between the "new" adaptation network on a few labeled data and "old" adaptation network on unlabeled data. If the gradients of the two networks have the same direction, the base network is updated in the current direction; otherwise, it is updated in the opposite direction. With feedback learning, the base network is consistently updated to generate better pseudo labels on unlabeled data, thereby improving the performance of adaptation network. As a result, we can narrow the data gap and learn rich knowledge in the circuit learning paradigm, even with a few labeled data.

Effective selection and utilization of unlabeled data are crucial for achieving good performance in animal face alignment. It is essential to avoid selecting hard negatives in training as finetuning on such examples (e.g., Spheniscidae) can decrease accuracy for all other animals, as demonstrated in Figure 1(b). Furthermore, augmenting labeled data with positive examples that have rather accurate pseudo-predictions can help mitigate data scarcity. However, obtaining these measurements requires knowing their ground truth landmarks, which contradicts the unlabeled data. Fortunately, we discovered that shifts in predicted landmarks between different models can indicate prediction accuracy. With this observation, we propose a novel positive example mining method to identify positives, semi-hard positives, and hard negatives from unlabeled data to facilitate our Meta-CSKT learning.

To sum up, this paper makes the following contributions:

- We observe knowledge sharing among animals, which provides a fundamental premise for Meta-CSKT, the first to leverage bi-directional cross-species knowledge transfer for data-scarce animal face alignment. Furthermore, Meta-CSKT can be generalized to the applications that have knowledge sharing across categories but lack high-quality labeled data.
- We propose a novel positive example mining method to effectively utilize unlabeled data. It is a crucial module for Meta-CSKT as it mitigates the scarcity of labeled data and leads to good performance.
- We conduct extensive experiments on three datasets to demonstrate the effectiveness of Meta-CSKT. It significantly outperforms state-of-the-art on the horse dataset and Japanese Macaque species, while achieving comparable results to state-of-the-art methods on large-scale labeled AnimalWeb (e.g., 18K), using only a few labeled images (e.g., 40).

2 RELATED WORK

2.1 Human face alignment

Many methods for human face alignment [15, 18, 28, 41] employ deep learning models with cascaded architecture to decompose a complex alignment problem into several manageable sub-problems, progressively improving the accuracy and reliability of the estimated landmarks. However, these methods are computationally expensive due to the demanding CNN architectures used throughout the cascade. Other approaches improve human face alignment by using recurrent models [16, 29], dense 3D model fitting [11, 43], or jointly learning auxiliary attributes and landmark detection [5, 42]. The Hourglass CNN architecture with a residual connection [2, 19, 32] has been widely used to produce landmark heatmaps. Unlike landmarks, heatmaps also estimate the certainty of landmarks. Some methods stress the semantic ambiguity of landmarks by discriminating meaningful motion from random motion induced by annotation noise [19], or estimating a more general probability distribution for heatmaps [3]. Recently, HRNet [27] has been proposed to extract high resolution (HR) representation for human face alignment, achieving state-of-the-art results. Specifically, HRNet includes paralleled high-to-low convolutions and multi-resolution fusion for cross-resolution information exchange. Despite remarkable progress made by modern deep architecture [8, 10] in human face alignment, they often rely on large-scale annotated datasets (e.g., 300W [24], AFLW [14], and Menpo [6]), which are often unsuitable for animal face alignment applications.

2.2 Animal face alignment

Despite the enormous benefits and direct impact on animal healthcare, animal face alignment is a relatively unexplored problem. Early work [38] uses a cascaded shape regressor with hand-craft features to estimate sheep facial landmarks by designing better hand-craft features. WarpingNet [23] transfers knowledge between human and horse data sources by warping a horse to have a more human-like shape so that the human face alignment model could easily adapt to the horse's appearance. However, this method focuses on aligning horse and sheep that share certain similarities with humans and does not tackle general animal face alignment. Ref. [30] addresses dog face detection and alignment by jointly learning cascaded regressors on a large-scale dog dataset. Ref. [9] introduces auxiliary head pose information to facilitate sheep face alignment and reduce the need for large-scale annotated data. Recently, a large-scale annotated animal dataset AnimalWeb [13] has been made publicly available. However, the accuracy is much lower than that of human face alignment when applying the popular human face alignment model [2, 35], indicating animal face alignment is challenging and largely unsolved. MDMD [7] uses the shared landmark semantic group prior to training two datasets that vary in landmark definitions and domains to enhance small dataset face alignment. DIFE [40] extracts common features shared across interspecies as dense face embedding, which can benefit various applications such as animal photo search. It first synthesizes pseudo pair images through the latent space exploration of StyleGAN2 [12] to find implicit associations between different animal faces and uses the semantic matching loss to combat extreme shape differences between species. POMNet [36] is proposed to predict the poses of any objects. However, it requires labeled support images during testing, which are not available in our application and thus cannot be used as a comparison method. In contrast, our Meta-CSKT only requires scarce annotations and is the first to leverage bi-directional cross-species knowledge transfer for animal face alignment based on knowledge sharing among animals.

2.3 Learning with unlabeled data

Annotating faces can be labor-intensive, error-prone, and difficult to maintain semantic consistency. Learning with unlabeled data can be a feasible way to this dilemma and some general methods such as pseudo label [17], noisy student [34] are proposed. Fixmatch [26] simplifies the learning process by training the model with high-confidence pseudo labels. UDA [33] improves semi-supervised learning by incorporating data augmentation [4] to limit the invariance of model predictions to input noise. MPL [20] enables the teacher network to adjust based on student's performance feedback on labeled data, which improves pseudo label [17]. CPGML [39] proposes inexactly supervised meta-learning to use coarse-grained labels of training samples to reduce the need for labeled data. In contrast, our Meta-CSKT applies semi-supervised method for animal face alignment, which is a regression task. This distinguishes it from the aforementioned methods designed for classification tasks.

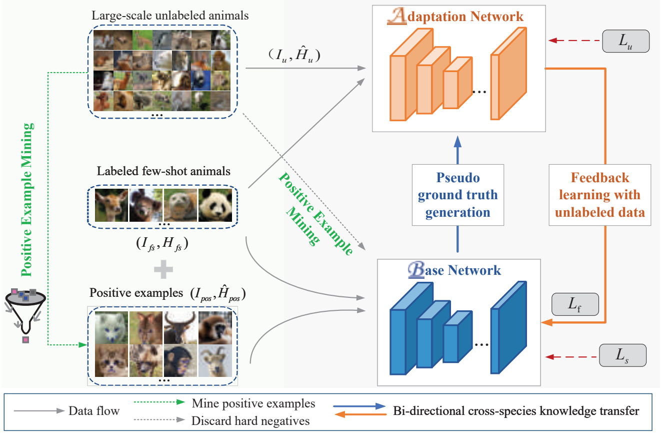


Figure 2: The Meta-CSKT training framework. The left part shows the application of positive example mining on large-scale unlabeled data, with positive examples augmenting the labeled data, and hard negatives being discarded. The right part shows the bi-directional cross-species knowledge transfer between the adaptation network and the base network.

3 METHOD: META-CSKT

3.1 Framework Overview

As shown in Figure 2, Meta-CSKT utilizes labeled few-shot animals for training, which consists of an adaptation network ($\mathcal{A}$) and a base network ($\mathcal{B}$). Both networks have the same network architecture with independent weights and are connected via bi-directional cross-species knowledge transfer. We use a pretrained face human alignment model to initialize their weights. Our positive example mining method is used to identify positives, semi-hard positives, and hard negatives in unlabeled data. Positive examples are used to augment the labeled data (as shown by the left dotted arrow), while hard negative examples are discarded from unlabeled data (as shown by the right dotted arrow) to avoid negatively impacting the performance.

At each generation, the adaptation network uses generated pseudo ground truth $\hat{H}_u$ (i.e., by applying the base network) on large-scale unlabeled animals I_u for training. The base network learns prior knowledge from labeled few-shot animals I_{fs} and the mined positive examples I_{pos}. With feedback learning, the base network is consistently updated to generate better pseudo heatmaps on unlabeled data, thereby improving the performance of the adaptation network.

Finally, the adaptation network outperforms the base network in the circuit learning paradigm. During the inference stage, only the adaptation network is used for animal face alignment.

3.2 Adaptation Network ($\mathcal{A}$)

The adaptation network utilizes large-scale unlabeled data to leverage its comprehensive intra- and inter-species variety. However, the hard-to-train negative examples can lead to bad local minima early in training, or even cause the animal model to collapse [25]. To address this issue, we introduce positive example mining $\mathtt{M}(\cdot)$ (in Section 3.4), which filters out hard negatives during training.

3.2.1 Pseudo ground truth generation ($\mathcal{B} \longrightarrow \mathcal{A}$). Animal face alignment aims to detect the locations of K landmarks that cover the major facial features around key face components (i.e., eyes, nose, and lips) from an image. Modern methods transform this problem by estimating K heatmaps, with each heatmap representing the location confidence of a landmark. The base network generates pseudo heatmaps $\hat{H}_u$ for unlabeled animal images I_u. To train the adaptation network, Meta-CSKT encourages two networks to predict similar heatmaps on unlabeled data using a loss $\mathcal{L}_u$:

$$\mathcal{L}_u = \left\|\hat{H}_u - \mathcal{A}(I_u;\theta_{\mathcal{A}})\right\|^2, \tag{1}$$

where $\hat{H}_u$ is the hard pseudo ground truth which can be derived from soft prediction $\mathcal{B}(I_u;\theta_{\mathcal{B}})$ in two steps: 1) extracting landmarks from soft prediction with highest confidence; 2) applying 2D Gaussian centered on each landmark location with a standard deviation of 1 pixel. Unlike ground truth labels, pseudo heatmaps change dynamically during training. In the meta-train phase, the parameters of the adaptation network $\theta_{\mathcal{A}}$ are updated. In the meta-test phase, feedback learning is introduced to enhance the base network and refine pseudo labels, further improving the performance of the adaptation network.

3.3 Base Network ($\mathcal{B}$)

The base network is trained using few-shot animal images I_{fs} along with online *mined* positive examples I_{pos}. Labeled few-shot animals have ground truth heatmaps but the positive examples do not. To overcome this challenging, the pseudo heatmaps of positive examples, predicted by the adaptation network, are used as ground truth. Positive example mining $\mathtt{M}(\cdot)$ ensures that only those with relatively accurate prediction are included as labeled training data. As the training progresses, the pseudo heatmaps of online positives gradually move towards true predictions, thereby strengthening the learning of the base network and largely reducing the need for labeled data. The base network is trained using a supervised loss ($\mathcal{L}_s$) and a feedback loss ($\mathcal{L}_f$):

$$\mathcal{L}_{\mathcal{B}} = \mathcal{L}_s + \mathcal{L}_f. \tag{2}$$

where the supervised loss is used for few-shot and positive data learning while the feedback loss is applied to mined unlabeled data.

3.3.1 Few-shot and positive data learning ($\mathcal{A} \longrightarrow \mathcal{B}$). For few-shot data, we utilized the ground truth heatmaps H_{fs} as targets. However, for mined positive data, we employ the pseudo label $\mathcal{A}(I_{pos};\theta_{\mathcal{A}})$ generated by the adaption network as targets, as the ground truth is not available. The base network is trained to minimize the mean square error between predicted heatmaps and their targets. For mined positive data, the pseudo label is generated by $\mathcal{A}(I_{pos};\theta_{\mathcal{A}})$ since the ground truth is not available. $\mathcal{L}_s$ is denoted as:

$$\mathcal{L}_s = \left\|H_{fs} - \mathcal{B}(I_{fs};\theta_{\mathcal{B}})\right\|^2 + \left\|\mathcal{A}(I_{pos};\theta_{\mathcal{A}}) - \mathcal{B}(I_{pos};\theta_{\mathcal{B}})\right\|^2, \tag{3}$$

where $\theta_{\mathcal{B}}$ are the parameters of the base network and the targets of mined positive examples are determined by the adaptation network. In this way, the cross-species knowledge learned by the adaptation network is implicitly transferred to the base network.

3.3.2 Feedback learning with mined unlabeled data ($\mathcal{A} \longrightarrow \mathcal{B}$). In Meta-CSKT, the intuition behind the base network update is the relationship between the "new" adaptation network on few-shot data and "old" adaptation network on unlabeled data. The adaptation network estimates the cognitive differences between few-shot and large-scale unlabeled data to update the base network as feedback. If the gradients of two networks have the same direction, the base network is updated in the current direction; otherwise, it is updated in the opposite direction. To achieve it, we formulate the feedback loss as the product of two terms, which are:

$$\mathcal{L}_f = f \cdot \left\|\hat{H}_u - \mathcal{B}(I_u;\theta_{\mathcal{B}})\right\|^2, \tag{4}$$

where the first term f is the feedback coefficient that determines the direction and strength of the update; the second term is the loss of the base network on unlabeled data. Specifically, the feedback coefficient f is defined as:

$$\begin{aligned} f = \eta_{\mathcal{A}} \cdot (\nabla_{\theta_{\mathcal{A}}^{(t+1)}} \mathrm{MSE}(H_{fs}, \mathcal{A}(I_{fs};\theta_{\mathcal{A}}^{(t+1)}))^{\top} \cdot \\ \nabla_{\theta_{\mathcal{A}}} \mathrm{MSE}(\hat{H}_u, \mathcal{A}(I_u;\theta_{\mathcal{A}}^{(t)}))), \end{aligned} \tag{5}$$

where MSE represents the mean square error loss $\|\cdot\|^2$ between the estimated and (pseudo) ground truth heatmaps. f is calculated as a dot product of two terms: the gradients of the "new" adaptation network on few-shot data and the gradients of the "old" adaptation network on large-scale unlabeled data. The sign of f will determine the direction of the update, while the absolute value of f will determine its strength. The adaptation network uses pseudo labeled data to update the parameters to $\mathcal{A}^{(t+1)}$. In particular, we approximate it with the parameters obtained from $\mathcal{A}^{(t)}$ by updating the base network parameters on $(I_u, \hat{H}_u)$, i.e., $\theta_{\mathcal{A}}^{(t+1)} = \theta_{\mathcal{A}}^{(t)} - \eta_{\mathcal{A}} \nabla_{\theta_{\mathcal{A}}} \mathrm{MSE}(\hat{H}_u, \mathcal{A}(I_u;\theta_{\mathcal{A}}))$.

3.4 Positive Example Mining

To represent how close the predicted heatmaps of two models (i.e., human and animal models) are to the ground truth for an unlabeled image, we use a triplet. The triplet captures the distance between data and the model to represent NME error. A smaller distance indicates a more accurate prediction. As illustrated in the upper part of Figure 3, our focus is on three types of triplets of unlabeled data. (a) *positives*: both the human and animal models generate similar and accurate heatmaps. (b) *semi-hard positives*: the animal model outperforms the human model by a large margin, but the predictions are less accurate compared to the positives. (c) *hard negatives*: both the human model and animal model fail to align animal images and cannot predict meaningful heatmaps. If the ground truth is known, it is easy to distinguish three types of unlabeled data. In practice, it is infeasible to directly generate such triplets as we do not have

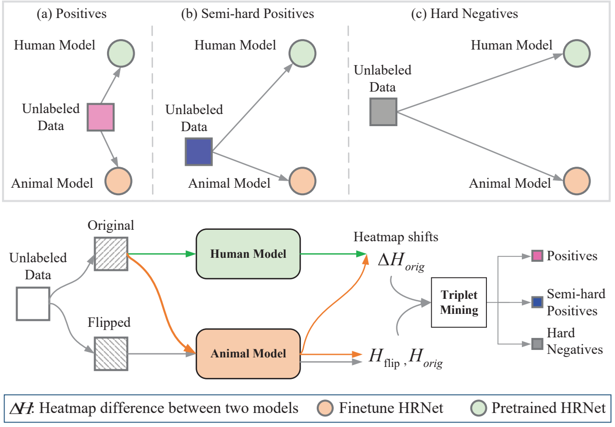


Figure 3: Positive example mining M(·) is used to identify three types of triplets of unlabeled data when no ground truth is available. A triplet is used to represent how close the predicted heatmaps of two models (i.e., human and animal models) are to the ground truth heatmaps.

access to ground truth heatmaps. When no ground truth is available, the task is accomplished through positive example mining.

Using hard negative examples can lead to poor training as mislabeled pseudo labels will dominate the learning of adaptation networks. We avoid this issue by excluding such hard negative data from training using the flip constraint. If the model can produce an accurate prediction on the original image, the predicted heatmaps of the original and flipped images should satisfy the flipped relationship, and vice versa. As a result, only positive and semi-positive examples are used as the unlabeled data. Unlike most methods that use feature extraction or loss variation for unlabeled data selection [25], we utilize the landmark shifts between pretrained and finetune models to determine the difficulty of aligning different animal species. We then use positive examples with reliable pseudo ground truth to augment labeled data and *online* update their pseudo labels. As illustrated in the lower part of Figure 3, the heatmap shifts between human and animal models of the original image are denoted as ΔH_{orig}. The predictions generated by applying the animal model to the original and flipped images are denoted as H_{orig} and H_{flip}, respectively. The process of triplet mining to identify positives, semi-hard positives, and hard negatives is as follows:

- If $\left\|H_{flip} - Flip(H_{orig})\right\|^2 > T_{neg}$, then the unlabeled data is hard negative. $Flip(\cdot)$ represents horizontal flip operation and T_{neg} is the threshold. (see Figure 3(c))
- If heatmap shifts $\Delta H_{orig} < T_{pos}$, then the unlabeled data is positive. T_{pos} is the threshold. (see Figure 3(a))
- If $\left\|H_{flip} - Flip(H_{orig})\right\|^2 \leqslant T_{neg}$ and heatmap shifts $\Delta H_{orig} \geqslant T_{pos}$, then the unlabeled data is semi-positive. (see Figure 3(b)).

3.5 Algorithm for Meta-CSKT

We listed detailed step-by-step pseudo-code for Meta-CSKT in Algorithm 1. Meta-CSKT extracts rich animal face alignment knowledge from large-scale unlabeled data via bi-directional cross-species knowledge transfer. At each generation, the adaptation network is first updated in line 7 by minimizing the unsupervised loss $\mathcal{L}_u$ on mined unlabeled data. This results in the transfer of knowledge from $\mathcal{B}$ to $\mathcal{A}$ via pseudo ground truth generation. The base network is then updated in line 16 using two losses: the supervised loss $\mathcal{L}_s$ and the feedback loss $\mathcal{L}_f$, to guide the learning process of the base network, which are illustrated in line 10 and line 14, respectively. We update the mined positive data online at intervals of s by using the adaptation network, which strengthens the learning of Meta-CSKT. As a result, knowledge transfers from $\mathcal{A}$ to $\mathcal{B}$ via few-shot and positive data learning and feedback learning with unlabeled data. By design, the two networks continuously complement each other via bi-directional cross-species knowledge transfer for animal face alignment and circumvent the data scarcity of labeled data.

Algorithm 1 Training procedure of Meta-CSKT

```
Input: Few-shot data $\mathcal{D}_{fs}$ and unlabeled data $\mathcal{D}_u^{'}$;
Human model $\mathcal{H}$; Mining thresholds: $\mathsf{T}_{\mathsf{pos}}, \mathsf{T}_{\mathsf{neg}}$; Mining interval: $s$
Outputs: $\Theta_{\mathcal{A}}^{(T)}$
Initialize: $\theta_{\mathcal{B}}^{(0)}$ and $\theta_{\mathcal{A}}^{(0)}$ with finetuned animal model $\mathcal{F}$
 1: Get mined positive and unlabeled data by using $\mathcal{H}$ and $\mathcal{F}$:
    $\mathcal{D}_{pos}, \mathcal{D}_u \leftarrow \mathsf{M}(\mathcal{D}_{\mathsf{u}}^{'}, \mathsf{T}_{\mathsf{pos}}, \mathsf{T}_{\mathsf{neg}})$
 2: for $t = 0 \ldots T - 1$ do
 3:     Get new labeled data $\mathcal{D}_l \leftarrow \mathcal{D}_{fs} \cup \mathcal{D}_{pos}$
 4:     $I_{fs}, H_{fs}, I_{pos}, \hat{H}_{pos} \leftarrow \text{SampleMiniBatch}(\mathcal{D}_l)$
 5:     $I_u \leftarrow \text{SampleMiniBatch}(\mathcal{D}_u)$
 6:     $\hat{H}_u \leftarrow \text{Forward}(I_u, \theta_{\mathcal{B}}^{(t)})$
 7:     Update the adaptation network using pseudo label:
 8:     $\theta_{\mathcal{A}}^{(t+1)} \leftarrow \theta_{\mathcal{A}}^{(t)} - \eta_{\mathcal{A}} \nabla_{\theta_{\mathcal{A}}} \text{MSE}(\hat{H}_u, \mathcal{A}(I_u; \theta_{\mathcal{A}}))$
 9:     Compute the base network's gradient on few-shot and
    mined positive data:
10:     $g_{\mathcal{B},s}^{(t)} \leftarrow \nabla_{\theta_{\mathcal{B}}} \text{MSE}(H_{fs}, \mathcal{B}(I_{fs}; \theta_{\mathcal{B}})) + \text{MSE}(\hat{H}_{pos}, \mathcal{B}(I_{pos}; \theta_{\mathcal{B}}))$
11:     Compute the base network's feedback coefficients:
12:     Applying Equation (5)
13:     Compute the base network's gradient via feedback:
14:     $g_{\mathcal{B},f}^{(t)} \leftarrow f \cdot \nabla_{\theta_{\mathcal{B}}} \text{MSE}(\hat{H}_u, \mathcal{B}(I_u; \theta_{\mathcal{B}}))$
15:     Update the base network:
16:     $\theta_{\mathcal{B}}^{(t+1)} \leftarrow \theta_{\mathcal{B}}^{(t)} - \eta_{\mathcal{B}} \cdot (g_{\mathcal{B},s}^{(t)} + g_{\mathcal{B},f}^{(t)})$
17:     for $t = s - 1, 2s - 1 \ldots$ do # update mined positive label
18:         $\hat{H}_{pos} \leftarrow \text{Forward}(I_{pos}, \theta_{\mathcal{A}}^{(t)})$
19:         Get new mined positive data: $\mathcal{D}_{pos} \leftarrow I_{pos}, \hat{H}_{pos}$
20:     end for
21: end for
22: return $\Theta_{\mathcal{A}}^{(T)}$
```

4 EXPERIMENTS

4.1 Datasets and Metrics

4.1.1 Datasets. We conduct our experiments on three widely used benchmark datasets for animal face alignment. **Horse Facial Keypoint dataset [23]** is an annotated horse dataset that contains 3717 horse images. Among them, 3531 training images and 186 testing images are annotated with 5 landmarks. **Japanese Macaque**

Species is a subset of AnimalWeb [13] and is used as an individual test benchmark in a recent work [7]. This dataset contains 133 Japanese Macaque images, including 100 training and 33 testing money faces. **AnimalWeb** is by far the most challenging and largest annotated animal face. It contains 22.4K annotated faces, offering 350 animal species with a variable number of animal faces in each species. The animal faces are annotated with 9 landmarks to cover major facial features around the eyes, nose, and lips. We pretrain our human face alignment model with the AFLW dataset [14].

4.1.2 Performance metrics. We use the Normalized Mean Error (NME) as a metric to calculate the Euclidean distance between the predicted and ground truth landmarks, which is then normalized by the face bounding box size. In addition to NME, we also report results using the failure rate as defined in [23]. If the Euclidean distance of landmarks is greater than 10% of the face size, it is considered a failure (referred to as failure@0.1(NME)). For evaluation, we also use the failure rate@0.08(NME) error as defined in [13]. *For both NME and failure rate, lower values indicate better performance.*

4.2 Implementation Details

4.2.1 Experimental protocols. For the Horse Facial Keypoint dataset, we follow the standard protocol [23] and 700 horse images from AnimalWeb are selected as unlabeled data. For the Japanese Macaque Species dataset, we follow the experimental setting in MDMD [7], where we use the first 100 images for training and the remaining 30 images for testing. To obtain the unlabeled data, we select primate animals (excluding the Japanese Macaque) from AnimalWeb based on biological taxonomy. The total number of unlabeled data is 3422. For AnimalWeb, we only use a few images as labeled data for training and the remaining images as unlabeled data. We evaluate our model under two settings: known species and unknown species. In the known species settings, we use the remaining images of the training species for testing. In the unknown species settings, we randomly select the same amount of unknown species as [13] from the remaining images for testing. We compare our method with [13], which follows similar settings, but with subtle differences as they require training with large-scale training data. The evaluation protocol is more stringent for our method, as the comparison methods require more than 17K training images, whereas we only use a small number of labeled images (e.g., 40). We conduct the ablation study on AnimalWeb using 40 labeled images. Effect of different few-shot animals is presented *in the supplementary.*

4.2.2 Training details. For positive example mining, we update positive examples at an interval of 600 steps. We set T_{pos} and T_{neg} to 0.05 and 0.2, respectively, resulting in 7342 positive images and 1822 hard negative images. Specifically, we select 7342 over 22,450 unlabeled examples to augment labeled data. During training, the number of positives consistently increases from 588 (@step 100) to a final 7342(@step 1300). Afterward, the model continues training with 7342 positive examples until convergence. Our method is relatively insensitive to different values of T_{pos} and T_{neg}, especially when the number of included positives and excluded negatives is large. We provide an analysis of their effects *in the supplementary.*

All input training images are cropped and resized to 255 × 255 pixels and the output heatmap has a size of 64 × 64 pixels. We use

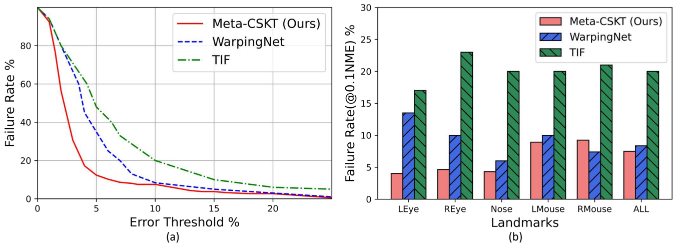


Figure 4: (a) Average landmark detection failure rate. (b) Failure rate@0.1NME comparison for 5 landmarks (i.e., left eye, right eye, nose tip, left mouth corner, and right mouth corner). 'ALL' is the average landmarks results.

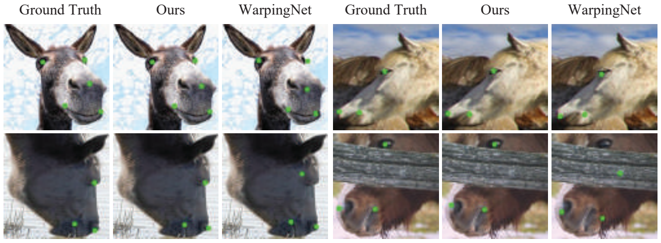


Figure 5: Examples of predicted landmarks for horses.

HRNet trained on the AFLW dataset as the backbone for base and adaptation networks by default. The learning rate is initialized by $1e-4$ and further decayed with a cosine annealing strategy. The batch size is set to 4 and 16 for labeled and unlabeled data. The entire training steps for Horse Facial Keypoint dataset, Japanese Macaque Species and AnimalWeb are 5,000, 500 and 10,000, respectively. It is trained end-to-end with one Nvidia Titan RTX GPU.

4.2.3 Comparison methods. We compare Meta-CSKT with eight state-of-the-art baselines, namely TIF, WarpingNet, MDMD, ViT-Pose, HG2-Known, HG2-Unknown, HG3-Known and HG3-Unknown on three datasets. Among them, WarpingNet, MDMD, and HG- models achieved the best performance on Horse Facial Keypoint dataset, Japanese Macaque Species and AnimalWeb, respectively. Nevertheless, our general solution applied to horse and Japanese Macaque alignment still significantly outperforms SOTAs (WarpingNet and MDMD). Moreover, we achieve comparable results with SOTAs on AnimalWeb using only a few labeled data (e.g., 40).

4.3 Results on Horse Facial Keypoint dataset

4.3.1 Quantitative results. We compare our Meta-CSKT with the baseline models on the Horse Facial Keypoint dataset. The results presented in Figure 4 show that our method significantly outperforms state-of-the-art methods. Figure 4(a) illustrates the average landmark detection failure rate , with a low value indicating better performance. Our method shows consistently superior accuracy across all thresholds, especially at low NME error, indicating that we can predict landmarks more accurately. Figure 4(b) shows the histogram of the average failure rate@0.1NME for landmark detection. Our method significantly reduces the failure rate, especially

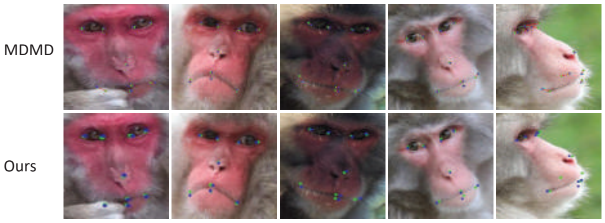


Figure 6: Results comparison between MDMD [7] and our method on Japanese Macaque. Results of MDMD are directly taken from their paper. The ground truth landmarks and predicted landmarks are indicated as green and blue.

Table 1: A comparison with some state-of-the-art methods on Japanese Macaque Species.

Methods	MDMD Base [7]	MDMD 300W [7]	ViTPose+B [37]	Ours
NME	3.66	3.44	4.69	2.96

for the left eyes and right eyes, regardless of challenges such as occlusion and large pose variations.

4.3.2 Qualitative results. We illustrate some qualitative examples of predicted landmarks by various methods in Figure 5. The results demonstrate that our method generates more precise landmarks than WarpingNet, which is consistent with the results in Figure 4. Particularly, in the first row, our method accurately predicts the mouth corner while WarpingNet does not. In the second row, WarpingNet predicts the eyes inaccurately due to severe occlusion and large pose, whereas our method is robust to such variations since Meta-CSKT is more successful in extracting discriminative features.

4.4 Results on Japanese Macaque Species

4.4.1 Quantitative results. We compare our Meta-CSKT with state-of-the-art methods including MDMD [7] and ViTPose [37] on the Japanese Macaque Species dataset. MDMD is trained on the Japanese Macaque Species as well as a large-scale human facial landmark dataset. The results reported for MDMD are reproduced from the original paper. ViTPose is a vision transformer proposed for human pose estimation, and therefore, cannot be directly compared to our paper. Hence, we train ViTPose under our experiment settings and conduct extensive hyperparameter tuning for performance optimization. The comparison results are shown in Table 1, where our method demonstrates significantly lower NME error, outperforming the compared methods by a large margin.

4.4.2 Qualitative results. We use the same set of illustration examples as MDMD [7] for fair comparison as their source code is not available. Figure 6 displays the predicted landmarks of different methods on Japanese Macaque, with the predicted results of MDMD taken from the paper. Our method demonstrates higher accuracy in predicting landmarks, which is consistent with the results in Table 1, particularly in cases of occlusion, such as mouth corner obscured by hands (the first example) and self-occluded eye regions caused by large pose variation (the last example).

Table 2: A comparison with some state-of-the-art animal face alignment models on AnimalWeb in known and unknown species settings. Results include error: NME/ FailureRate@0.1(NME)/ FailureRate@0.08(NME). All compared methods are trained with large-scale labeled data.

Models	# Labeled Img	NME/FR@0.1/FR@0.08
HG2-Known	17.96K	5.22%/–/16.4%
HG3-Known	17.96K	5.12%/–/16.3%
Ours-Known	40	5.61%/11.2%/18.5%
Ours-Known	80	5.55%/12.7%/19.6%
HG2-Unknown	17.62K	6.14%/–/22.0%
HG3-Unknown	17.62K	5.96%/–/20.7%
Ours-Unknown	40	7.44%/20.9%/28.2%
Ours-Unknown	80	7.21%/21.0%/27.9%

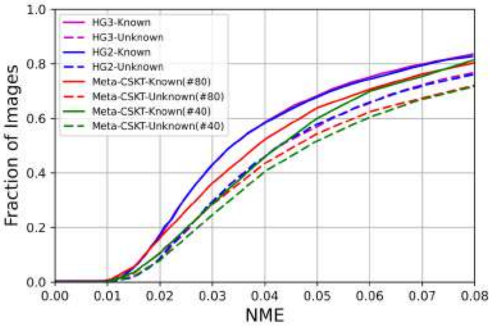


Figure 7: Comparison between state-of-the-arts using large-scale labeled data and Meta-CSKT with a few labeled data.

4.5 Results on AnimalWeb

We compare our Meta-CSKT with state-of-the-art baseline models on the AnimalWeb dataset. These comparative models use *large-scale labeled* AnimalWeb for training. Table 2 shows a performance comparison on AnimalWeb in terms of two settings. For the known species settings, Meta-CSKT can achieve comparable results using only a few labeled images. For example, Meta-CSKT trained with 80 labeled images achieves 5.55% error, while HG2 with 17.96K labeled images is 5.22% error. For the unknown species settings, we observe a small gap (NME difference). For example, Meta-CSKT performs about 1 unit worse than HG2. The reason may be that a few labeled images in Meta-CSKT are weakly related to the tested rare animal species, hindering the learning of cross-species knowledge transfer. Fortunately, we observe a trend that training with more labeled images improves accuracy. For instance, Meta-CSKT trained with 80 labeled images yields lower NME errors than using 40 labeled images. In addition, we also observe a similar trend in the Cumulative Error Distribution (CED) curves in Figure 7. Higher values are better. A small gap exists between our method and the HG- models trained with large-scale data. However, this gap significantly narrows for smaller NME thresholds. For instance, at 0.01 NME, there is almost no difference between these two methods under both settings, which verifies the effectiveness of our method.

4.6 Ablation Study

4.6.1 Effect of Meta-CSKT design. Table 3 presents the performance of Meta-CSKT using different losses and positive example mining on AnimalWeb. The results show that (1) Using three losses together, model 3 achieves the best performance compared to the first two models, indicating the effectiveness of bi-directional cross-species knowledge transfer. (2) The unsupervised loss $\mathcal{L}_u$ enhances animal face alignment by a large margin, verifying the effectiveness of pseudo ground truth generation. (3) Incorporating the feedback loss $\mathcal{L}_f$ consistently updates the base network to generate better pseudo heatmaps on unlabeled data, further improving the results. (4) Avoiding selecting hard negative examples from unlabeled data during training is beneficial, as model4 further improves model3. (5) Table 3 clearly shows that including positive examples (i.e., with their pseudo label) as labeled data significantly improves the performance of Meta-CSKT. This strengthens the learning of the base network as their pseudo labels are continuously refined during training, leading to the best performance of Meta-CSKT.

Table 3: Known/Unknown species NME errors (%) of Meta-CSKT on AnimalWeb with different loss functions (i.e., $\mathcal{L}_s$, $\mathcal{L}_u$, $\mathcal{L}_f$) and our positive example mining (i.e., excluding hard negatives, and including positive examples as labeled data).

Models	Loss			Positive Example Mining		NME Error
	$\mathcal{L}_s$	$\mathcal{L}_u$	$\mathcal{L}_f$	Exc. Negative	Inc. Positive	
1	✓	✗	✗	✗	✗	6.67%/10.07%
2	✓	✓	✗	✗	✗	6.03%/9.03%
3	✓	✓	✓	✗	✗	5.93%/8.88%
4	✓	✓	✓	✓	✗	5.89%/8.84%
Ours	✓	✓	✓	✓	✓	5.61%/7.44%

4.6.2 Insights of positive example mining. We further show the mined animal family distribution in Figure 8. The x-axis is the family index, sorted in ascending order of NME by the human model. This means that from left to right, the animal species become increasingly challenged. The y-axis indicates the ratio of selected examples from unlabeled data. The results show that most excluded hard negatives are located in the tail of the histogram, and most included positives are located at the head of the histogram, which is consistent with our observation in Figure 1. Furthermore, for negative examples, only a few animals that are far from the tail of the histogram are mined because such animals are challenging and they fail to pass our flipping constraints. On the other hand, for positive examples, some animals in the middle of the histogram are also mined, which is beneficial for general animal face alignment. This can facilitate bi-directional cross-species transfer as these animals are generally positively correlated with many other animals.

4.6.3 Visual analysis. We illustrate some qualitative examples of predicted landmarks generated by different baselines in Figure 9, which are in line with quantitative results in Table 3. From left to right, the accuracy of predicting landmarks gradually improves with loss design and positive example mining. As shown in Figure 9, the animal faces are from rare species of AnimalWeb and tend to exhibit large variations in pose, appearance, and expressions. Notably, despite using only a few labeled data during training, our Meta-CSKT produces accurate landmarks for most animal faces.

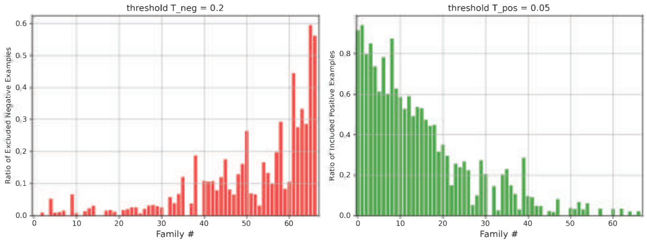


Figure 8: An illustration of mined animal family distribution by applying positive example mining.

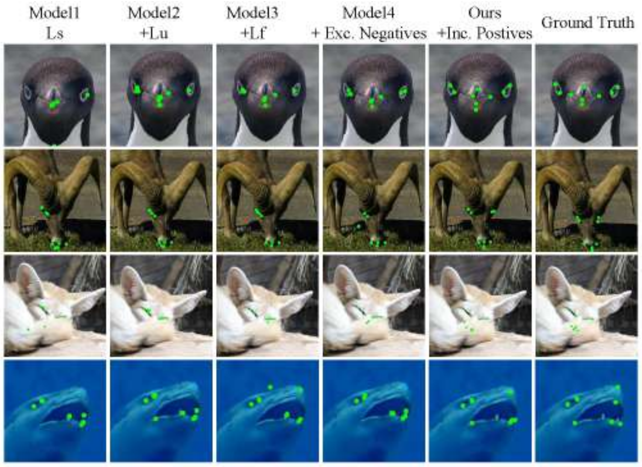


Figure 9: Qualitative examples comparing our Meta-CSKT and baselines on Animalweb of unknown species settings. Compared to the left column, the Baseline model in the right column adds a new loss or mining operation by using '+'.

5 CONCLUSION

In this paper, we propose the Meta-CSKT for data-scarce animal face alignment through meta optimization. Our method is motivated by two observations: i) the knowledge sharing among animals, and ii) the predicted accuracy revealed by landmark shifts between human and animal models. The first observation motivates bi-directional CSKT between labeled few-shot and large-scale unlabeled animals. The second observation motivates positive example mining to mitigate the data scarcity of labeled data. Extensive experiments on three datasets demonstrate the superiority of our method for animal face alignment with only a few labeled images.

ACKNOWLEDGMENTS

This work was partially supported by the National Natural Science Foundation of China No.62206123, Shenzhen Fundamental Research Program (Grant No. 20220815112848002) and the Guangdong Provincial Key Laboratory (Grant No. 2020B121201001).

REFERENCES

[1] Gaddi Blumrosen, David Hawellek, and Bijan Pesaran. 2017. Towards automated recognition of facial expressions in animal models. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2810–2819.

[2] Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*. 1021–1030.

[3] Lisha Chen, Hui Su, and Qiang Ji. 2019. Face alignment with kernel density deep neural network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6992–7002.

[4] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 702–703.

[5] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5203–5212.

[6] Jiankang Deng, Anastasios Roussos, Grigorios Chrysos, Evangelos Ververas, Irene Kotsia, Jie Shen, and Stefanos Zafeiriou. 2019. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *International Journal of Computer Vision* 127, 6 (2019), 599–624.

[7] David Ferman and Gaurav Bharaj. 2022. Multi-domain Multi-definition Landmark Localization for Small Datasets. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*. Springer, 646–663.

[8] Ivan Gogić, Jörgen Ahlberg, and Igor S Pandžić. 2021. Regression-based methods for face alignment: A survey. *Signal Processing* 178 (2021), 107755.

[9] Charlie Hewitt and Marwa Mahmoud. 2019. Pose-informed face alignment for extreme head pose variations in animals. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–6.

[10] Xin Jin and Xiaoyang Tan. 2017. Face alignment in-the-wild: A survey. *Computer Vision and Image Understanding* 162 (2017), 1–22.

[11] Amin Jourabloo and Xiaoming Liu. 2016. Large-pose face alignment via CNN-based dense 3D model fitting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4188–4196.

[12] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.

[13] Muhammad Haris Khan, John McDonagh, Salman Khan, Muhammad Shahabuddin, Aditya Arora, Fahad Shahbaz Khan, Ling Shao, and Georgios Tzimiropoulos. 2020. Animalweb: A large-scale hierarchical dataset of annotated animal faces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6939–6948.

[14] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. 2011. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE International Conference on Computer Vision Workshops*. IEEE, 2144–2151.

[15] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. 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.

[16] Hanjiang Lai, Shengtao Xiao, Yan Pan, Zhen Cui, Jiashi Feng, Chunyan Xu, Jian Yin, and Shuicheng Yan. 2016. Deep recurrent regression for facial landmark detection. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 5 (2016), 1144–1157.

[17] Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, Vol. 3. 896.

[18] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. 2017. Learning deep sharable and structural detectors for face alignment. *IEEE Transactions on Image Processing* 26, 4 (2017), 1666–1678.

[19] Zhiwei Liu, Xiangyu Zhu, Guosheng Hu, Haiyun Guo, Ming Tang, Zhen Lei, Neil M Robertson, and Jinqiao Wang. 2019. Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3467–3476.

[20] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. 2021. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11557–11568.

[21] Maheen Rashid. 2021. *Towards Automatic Visual Recognition of Horse Pain*. Ph.D. Dissertation. University of California, Davis.

[22] Maheen Rashid, Sofia Broomé, Katrina Ask, Elin Hernlund, Pia Haubro Andersen, Hedvig Kjellström, and Yong Jae Lee. 2022. Equine Pain Behavior Classification via Self-Supervised Disentangled Pose Representation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1646–1656.

[23] Maheen Rashid, Xiuye Gu, and Yong Jae Lee. 2017. Interspecies knowledge transfer for facial keypoint detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6894–6903.

[24] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 2013. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 397–403.

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[26] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems* 33 (2020), 596–608.

[27] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 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.

[28] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2013. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3476–3483.

[29] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. 2016. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4177–4187.

[30] Alzbeta Vlachynska, Zuzana Kominkova Oplatkova, and Tomas Turecek. 2018. Dogface detection and localization of dogface's landmarks. In *Computer Science On-line Conference*. Springer, 465–476.

[31] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 10 (2020), 3349–3364.

[32] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. 2018. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2129–2138.

[33] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems* 33 (2020), 6256–6268.

[34] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10687–10698.

[35] Pengfei Xiong, Guoqing Li, and Yuhang Sun. 2017. Combining local and global features for 3D face tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2529–2536.

[36] Lumin Xu, Sheng Jin, Wang Zeng, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. 2022. Pose for Everything: Towards Category-Agnostic Pose Estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*. Springer, 398–416.

[37] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. In *Advances in Neural Information Processing Systems*.

[38] Heng Yang, Renqiao Zhang, and Peter Robinson. 2016. Human and sheep facial landmarks localisation by triplet interpolated features. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–8.

[39] Jinhai Yang, Hua Yang, and Lin Chen. 2020. Coarse-to-Fine Pseudo-Labeling Guided Meta-Learning for Few-Shot Classification. *Technical report* (2020).

[40] Sejong Yang, Subin Jeon, Seonghyeon Nam, and Seon Joo Kim. 2022. Dense Interspecies Face Embedding. *Advances in Neural Information Processing Systems* 35 (2022), 33275–33288.

[41] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. 2014. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European Conference on Computer Vision*. Springer, 1–16.

[42] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[43] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. 2016. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 146–155.

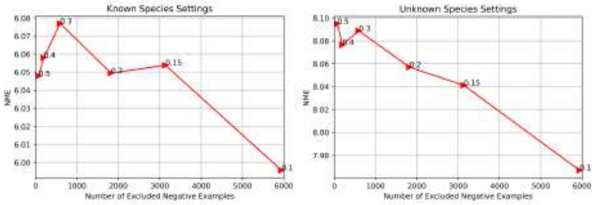


Figure 11: NME error of Meta-CSKT on AnimalWeb with varying numbers of excluded hard negative examples in known/unknown species settings.

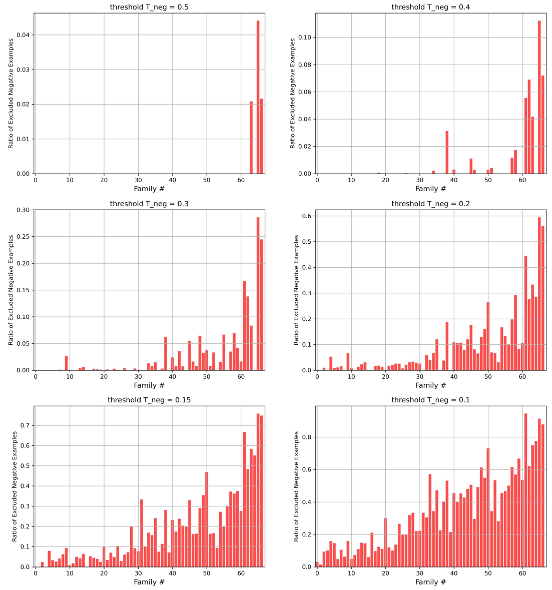


Figure 12: The animal family distribution of excluded hard negative examples by different threshold T_{neg}.

A EFFECT OF DIFFERENT LABELED FEW-SHOT ANIMALS

We report the performance of Meta-CSKT using different few-shot animals in Figure 10. We test different models on the rest of all training images. We explore the effect of few-shot animals on animal face alignment by varying numbers and species. With a pretrained human face alignment model, species are broadly divided into easy, medium, and hard species classes based on NME errors. Generally, images of the easy species class are similar in shape or appearance to humans, images of the medium/hard species classes are related/unrelated to humans. Figure 10 shows that increasing the number of images generally leads to better performance, except for the easy species class. For easy species class, the improvement is not significant because these images can already be well aligned by using the pretrained human model, and using more of these images cannot facilitate knowledge transfer across species. Moreover, using images from medium species class yields the best performance as they are more likely to transfer knowledge across species in either easy or hard classes. Thus, Meta-CSKT uses images from the medium species class as few-shot labeled animals.

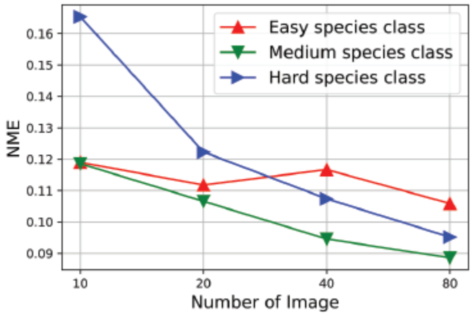


Figure 10: An illustration of the effect of few-shot labeled data in terms of quantity (e.g., 10,20,40,80) and animal species.

B ANALYSIS OF THRESHOLD T_{neg}

Figure 11 uses NME error of Meta-CSKT model as a function of the number of excluded negative examples on both settings. The thresholds T_{neg}, ranging from 0.5 to 0.1, are marked in the figure. Lower T_{neg} means more hard negative examples are excluded from the large-scale unlabeled data during training. The results show that (1) Meta-CSKT generally obtains better alignment accuracy when excluding large amount of hard negative examples (e.g., 6000, T_{neg} is 0.1) from the large-scale unlabeled data. (2) The performance slightly fluctuates when only a small amount of data is excluded, which is reasonable because the remaining negative examples hinder semi-supervised learning and knowledge transfer between different species in Meta-CSKT.

We further show the family distribution of the selected hard negative examples in Figure 12. The results show that most species of hard negatives are located in the tail of the histogram. As T_{neg} decreases, some animal species in the middle of the histogram are also selected. The reason is that some animals have large intra- and inter-species variations and cannot satisfy the flipping constraints.

C ANALYSIS OF THRESHOLD T_{pos}

Figure 13 uses Meta-CSKT model NME error as a function of the number of included positive examples on known and unknown species settings. The thresholds T_{pos}, ranging from 0.01 to 0.07, are marked in the figure. As results show, (1) including more positive examples as labeled data leads to better accuracy in both settings. (2) The accuracy drops when only a small number of positive examples (e.g., 944, T_{pos} is 0.02) are included. The decline is because most positive examples are human-like species that can be perfectly aligned with the human model and Meta-CSKT cannot benefit from bi-directional cross-species knowledge transfer.

We further show the family distribution of the selected positive examples in Figure 14. The results show that most species are located at the head of the histogram. As T_{pos} increases, more unlabeled data are mined as Meta-CKST positive examples. Specifically, some animal species in the middle of the histogram are selected. They can facilitate bi-directional cross-species transfer as these animal species are generally positively correlated with many other animals and knowledge is shared among them.

Table 4: Model complexity and running time comparison between our method and state-of-the-art methods.

Methods	Backbone	# Parameters	Inference time(s)
WarpingNet	WarpingNet	NA	NA
MDMD	ViT	86M	NA
ViTPose+B	ViT	86M	0.015s
HG2-Known/Unknown	HG-2	8.4M	0.045s
HG3-Known/Unknown	HG-3	12.5M	0.068s
Meta-CSKT(Our)	HRNetV2-W18	9.3M	0.003s

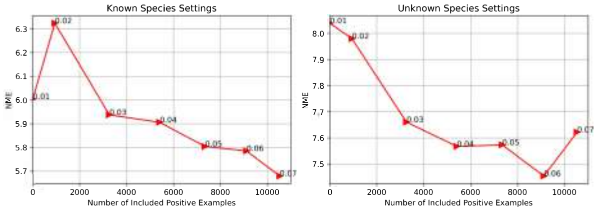


Figure 13: NME error of Meta-CSKT on AnimalWeb with varying numbers of included positive examples in known/unknown species settings.

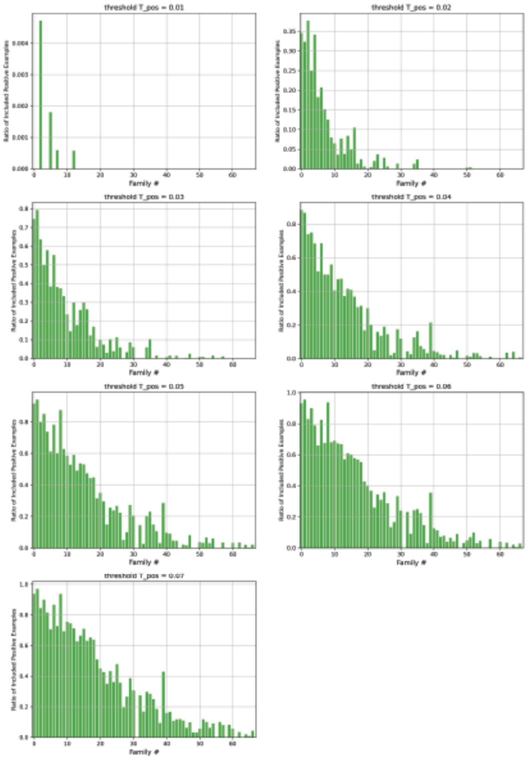


Figure 14: The animal family distribution of included positive examples by different threshold T_{pos}.

D MODEL COMPLEXITY AND RUNNING TIME COMPARISON

We compare our method with seven representative models: WarpingNet, MDMD, ViTPose-B, HG2(3)-Known/Unknown models. The comparison results are summarized in the Table 4. The results reported by WarpingNet and MDMD are from their original papers, as their code hasn’t been made publicly available. As shown, our method demonstrates competitive performance; it is the second lightweight model with 9.3M parameters and displays the fastest inference time of 0.003 seconds per image. We consciously chose not to include TIF in our comparison, as it’s a shallow method that learns with hand-crafted features.

E COMPARING WITH REPRESENTATIVE SEMI-SUPERVISED LEARNING METHODS METHODS

We compare our method with four state-of-the-art semi-supervised learning methods, including Pseudo leabel[r1], Unsupervised Data Augmentation(UDA)[r2], FixMatch[r3], and ScarceNet[r4] on the AnimalWeb dataset in Table 5. In our evaluation, Pseudo label is the most classic semi-supervised learning method. Both UDA and FixMatch methods incorporate pseudo labeling and consistency regularization with strong augmentations, which achieves better performance than Pseudo label. For UDA and FixMatch, we reimplement these semi-supervised approaches according to the open repository since they only show results for classification tasks. To ensure a fair comparison, we use the same human face alignment model as their pretrained model. ScarceNet focuses on animal pose estimation with scarce annotations, making it highly relevant to our work. We train ScarceNet under our experiment settings (i.e., using the same 40 few-shot labeled data and large-scaled unlabeled data) and conduct extensive hyperparameter tuning for performance optimization. The comparison results show that our Meta-CSKT outperforms other semi-supervised methods by a large margin.

Table 5: A comparison with some semi-supervised learning methods for animal face alignment.

Methods	Pseudo label	UDA	FixMatch	ScarceNet	Ours
NME(Known species)	6.03%	5.8%	5.94%	5.81%	5.61%
NME(Unknown species)	9.03%	7.48%	7.49%	8.32%	7.44%

REFERENCES OF APPENDIX

[r1] Lee, D.H., 2013, June. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Workshop on challenges in representation learning, ICML (Vol. 3, No. 2, p. 896).

[r2] Xie, Q., Dai, Z., Hovy, E., Luong, T. and Le, Q., 2020. Unsupervised data augmentation for consistency training. Advances in neural information processing systems, 33, pp.6256-6268.

[r3] Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A. and Li, C.L., 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems, 33, pp.596-608.

[r4]Li, C. and Lee, G.H., 2023. ScarceNet: Animal Pose Estimation with Scarce Annotations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17174-17183).