UniAP: Towards Universal Animal Perception in Vision via Few-Shot Learning

Meiqi Sun[1*], Zhonghan Zhao[2*], Wenhao Chai[3*], Hanjun Luo[1], Shidong Cao[1], Yanting Zhang[4], Jenq-Neng Hwang[3], Gaoang Wang[1,2,5†]

[1]Zhejiang University-University of Illinois Urbana Champaign Institute, Zhejiang University
[2]College of Computer Science and Technology, Zhejiang University
[3]Electrical and Computer Engineering Department, University of Washington
[4]Department of Computer Science and Technology, Donghua University
[5]Shanghai Artificial Intelligence Laboratory
gaoangwang@intl.zju.edu.cn

Abstract

Animal visual perception is an important technique for automatically monitoring animal health, understanding animal behaviors, and assisting animal-related research. However, it is challenging to design a deep learning-based perception model that can freely adapt to different animals across various perception tasks, due to the varying poses of a large diversity of animals, lacking data on rare species, and the semantic inconsistency of different tasks. We introduce UniAP, a novel Universal Animal Perception model that leverages few-shot learning to enable cross-species perception among various visual tasks. Our proposed model takes support images and labels as prompt guidance for a query image. Images and labels are processed through a Transformer-based encoder and a lightweight label encoder, respectively. Then a matching module is designed for aggregating information between prompt guidance and the query image, followed by a multi-head label decoder to generate outputs for various tasks. By capitalizing on the shared visual characteristics among different animals and tasks, UniAP enables the transfer of knowledge from well-studied species to those with limited labeled data or even unseen species. We demonstrate the effectiveness of UniAP through comprehensive experiments in pose estimation, segmentation, and classification tasks on diverse animal species, showcasing its ability to generalize and adapt to new classes with minimal labeled examples.

Introduction

Animal perception plays an essential role in learning and understanding animal behaviors (Ashwood, Jha, and Pillow 2022; Anderson and Donath 1990; Butail et al. 2015). Animal perception techniques have been employed in many visual tasks, such as pose estimation (Li and Lee 2023; Graving et al. 2019), pose tracking (Pereira et al. 2022; Lauer et al. 2022), face recognition (Shi et al. 2020), and semantic segmentation (Li et al. 2021). Existing animal perception models usually focus on specific animal species on specific tasks. To facilitate analyzing animal behaviors and assist animal-related research, it is highly demanded to design a universal animal perception model for diverse species and poses on various tasks.

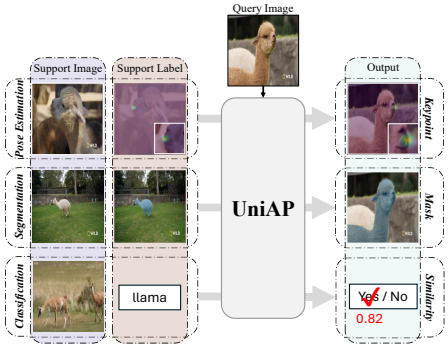

Figure 1: UniAP unifies animal pose estimation (top), segmentation (middle), and classification (bottom) tasks under a single model via few-shot learning. We use different support images of alpaca in Animal Kingdom dataset (Ng et al. 2022) as examples for different tasks.

Due to various animal species with different poses and semantic inconsistency of various visual tasks, it is challenging to build a universal animal perception model. Firstly, the variety of animal species requires the perception model to have a high generalization capability. The mainstream approaches for animal perception rely on large-scale, manually labeled data for supervised training, which is inefficient and suffers from annotation quality issues. Adapting the model to rare animals with only a few data examples or unseen species is difficult. Secondly, due to the semantic inconsistency of various tasks, it remains difficult to integrate information across different animal-related tasks and species to create a comprehensive representation of animal perception. This lack of unified cross-task models in multi-modal animal perception presents a significant challenge.

We claim that a universal few-shot model of animal perception tasks requires meeting certain criteria. Firstly, the model should have a unified architecture that can handle multiple tasks and share parameters across tasks to acquire generalizable knowledge. Secondly, the model ought to have the ability to work effectively with various animal species,

*These authors contributed equally.
†Corresponding author.

even in situations with limited data, allowing it to use few-shot prompts and yield superior outcomes across a wider range of species.

We propose **UniAP**, a Universal Animal Perception model that leverages few-shot learning to enable cross-species perception among various visual tasks, as shown in Fig. 1. Our proposed model takes prompt images and labels from different modalities, such as poses and masks, as the support set. With prompt guidance, UniAP can generate multi-task outputs for a query image. Specifically, images and labels are processed through a Transformer-based encoder with task-specific learnable bias parameters and a lightweight label encoder. Then a matching module is designed for aggregating information between prompt guidance and the query image, followed by a multi-head label decoder to generate outputs for various tasks. By capitalizing on the shared visual characteristics among different animals and tasks, UniAP enables knowledge transfer from well-studied species to those with limited labeled data or unseen species. We demonstrate the effectiveness of UniAP through comprehensive experiments in pose estimation, segmentation, and classification tasks on diverse animal species, showcasing its ability to generalize and adapt to new classes with minimal labeled examples.

Our core contributions are summarized as follows:

- We propose UniAP, a novel universal model that tackles the problem of unified learning of animal perception tasks: pose estimation, semantic segmentation, and classification.
- UniAP can take images and labels from various modalities as prompt guidance to generate outputs for rare species or even unseen animals with a few examples.
- Extensive experiments and ablation studies on various tasks and datasets demonstrate the effectiveness of UniAP.

Related Works

Animal Pose Estimation

The pose estimation task has achieved significant success in humans (Andriluka et al. 2014; Munea et al. 2020; Dang et al. 2019) or vehicles (Fang and López 2019; López, Agudo, and Moreno-Noguer 2019). There have been some prior works showing interest in pose estimation for animals (Pereira et al. 2019; Graving et al. 2019; Lauer et al. 2022; Cao et al. 2019; Li and Lee 2021; Zhang et al. 2022; Jiang et al. 2022; Shooter, Malleson, and Hilton 2021). Animal pose estimation plays an essential role in learning and understanding animal behavior (Anderson and Donath 1990; Butail et al. 2015). Recently, POMNet (Xu et al. 2022) aims to create a pose estimation model capable of detecting the pose of any class of object given only a few samples with keypoint definition. ScarceNet (Li and Lee 2023) addresses animal pose estimation with limited labeled data and unlabeled images. CLAMP (Zhang et al. 2023) attempts to bridge the gap by adapting the text prompts to the animal keypoints. FSKD (Lu and Koniusz 2022) is the first attempt to model keypoint detection as few-shot learning. There are also several benchmarks like APT-36k (Yang et al. 2022) and AP-10k (Yu et al. 2021) facilitating the research in animal pose estimation.

Animal Semantic Segmentation and Classification

Semantic segmentation (Guo et al. 2018; Wang et al. 2018) is a fundamental task in computer vision that assigns a class label to each pixel in an image. Recently, some works extend it to open-vocabulary (Luo et al. 2023; Ma et al. 2022), *i.e.*, segmenting objects from any categories by their textual names or description. SAM (Kirillov et al. 2023) aims to return a valid segmentation mask given any prompt, which opens up the possibility of zero-shot general object segmentation. Some works also focus on animal semantic segmentation (Li et al. 2021). But there is no specific benchmark for animal semantic segmentation. As for animal classification, some widely used benchmarks like MSCOCO (Lin et al. 2014) and ImageNet (Russakovsky et al. 2015) have plenty of images for animal classification.

Few-shot Learning in Computer Vision

Few-shot learning is a fundamental paradigm in computer vision that carries the promise of alleviating the need for exhaustively labeled data and has been widely explored in vision tasks such as semantic segmentation (Shaban et al. 2017; Iqbal, Safarov, and Bang 2022; Fan et al. 2022), instance segmentation (Michaelis et al. 2018; Fan et al. 2020b), and object detection (Fan et al. 2020a; Kang et al. 2019). Recently, VTM (Kim et al. 2023) proposes a universal few-shot learner for arbitrary dense prediction tasks. It utilizes non-parametric matching on embedded tokens of images and labels at the patch level, encompassing all tasks.

Methodology

We propose **UniAP**, a Universal Animal Perception framework, which is designed to flexibly adapt to diverse animals across domains with few-shot examples for multiple unified tasks. The overall architecture of UniAP is shown in Fig. 2. We first present the overview of our proposed UniAP, and then introduce the architecture of the model, followed by the training and inference of the framework.

Overview

We develop a unified few-shot animal perception model denoted as $\mathcal{F}$, which can provide predictions $\hat{Y}^q$ on pose estimation (PE), semantic segmentation (SS) and classification (CLS) for an unseen image (*query*) X^q given only a few labeled examples (*prompt set*) $\mathcal{P}$ for any given task in $\mathcal{T} = \{PE, SS, CLS\}$:

$$\hat{Y}^q = \mathcal{F}(X^q; \mathcal{P}), \quad \mathcal{P} = \{(X_i^p, Y_i^p)\}_{i \leq N}, \tag{1}$$

where (X_i^p, Y_i^p) represents an image-label pair in the prompt set, and N is the size of the prompt set.

We aim to find a universal function form $\mathcal{F}$ for Eq. (1) that can generate organized labels for any task within a unified framework. The query label is obtained by combining

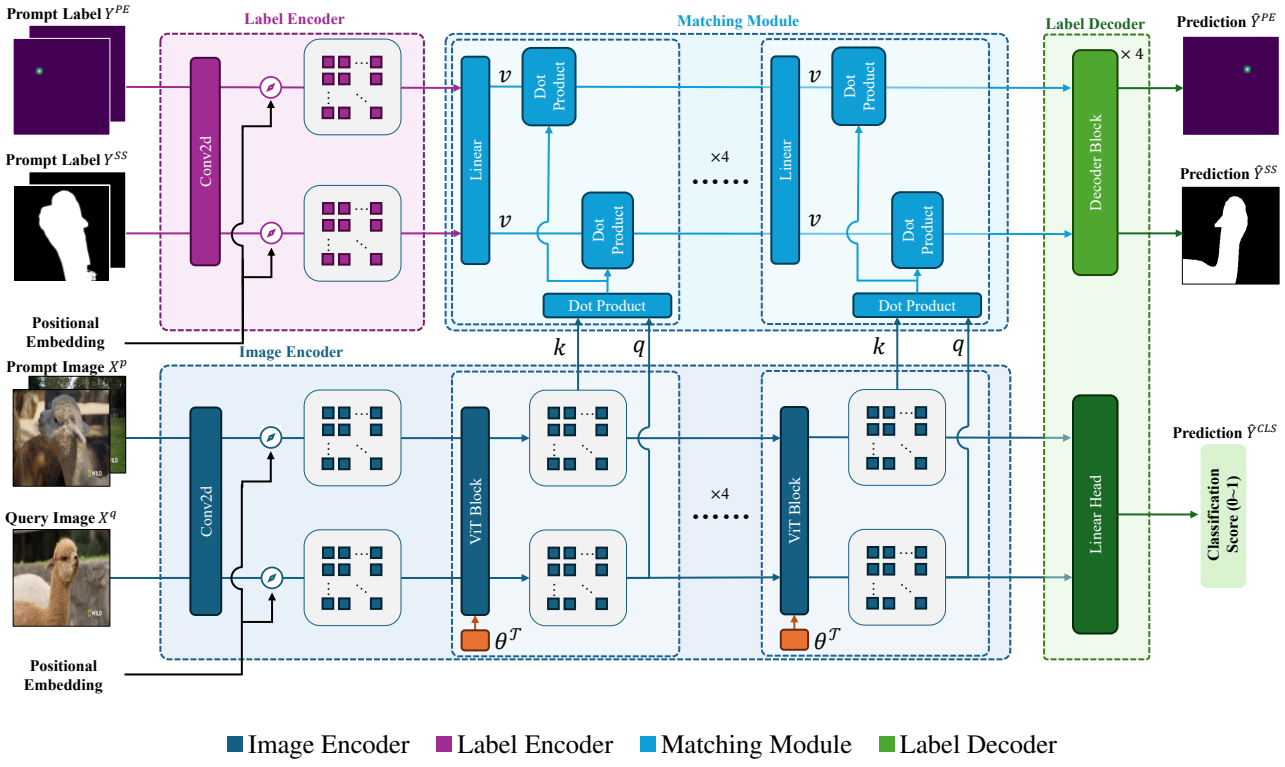

Figure 2: Overall architecture of UniAP. Our model is a hierarchical encoder-decoder with Query Image X^q given only a few labeled examples Prompt Set $\mathcal{P} = \{(X^p, Y^p)\}$. For any given task: pose estimation (PE) or semantic segmentation (SS), specific Prompt Images X^p correspond to Prompt Label Y^{PE} or Y^{SS}. Note that UniAP uses the feature pyramid network (FPN) to maintain the Affinity Matrix through multi-head attention layers in each hierarchy's matching module.

the prompt labels in a calculated manner. Given a query image X^q and a prompt set $\{(X_i^p, Y_i^p)\}_{i \leq N}$ with image-label pairs, the query label Y^q can be obtained as follows,

$$\mathcal{F}\left(X^q; \mathcal{P}\right) = h\left(\mathcal{M}\left(f_{\mathcal{T}}\left(X^q\right), \{f_{\mathcal{T}}\left(X_i^p\right)\}_{i \leq N}, \{g\left(Y_i^p\right)\}_{i \leq N}\right)\right), \tag{2}$$

where $f_{\mathcal{T}}$, g, $\mathcal{M}$ and h are image encoder, label encoder, matching module, and label decoder, respectively.

Specifically, prompt images $\{X_i^p\}_{i \leq N}$ and the query image X^q are encoded by the image encoder $f_{\mathcal{T}}$. Note that $f_{\mathcal{T}}$ contains task-specific lightweight trainable parameters $\theta_{\mathcal{T}}$. In our approach, we incorporate adaptability with task-specific parameters $\theta_{\mathcal{T}}$, which is crucial in reflecting the unique features of each task. Simultaneously, prompt labels $\{Y_i^p\}_{i \leq N}$ are encoded by a lightweight label encoder g, where g is shared across different tasks. After encoding, the embedded features of prompt set and the query image are aggregated in the matching module $\mathcal{M}$. The matching module $\mathcal{M}$ is formulated as follows,

$$\mathcal{M}\left(f_{\mathcal{T}}\left(X^q\right), \{f_{\mathcal{T}}\left(X_i^p\right)\}_{i \leq N}, \{g\left(Y_i^p\right)\}_{i \leq N}\right) = \sum_{i \leq N} \sigma\left(f_{\mathcal{T}}(X^q), f_{\mathcal{T}}(X_i^p)\right) \cdot g(Y_i^p), \tag{3}$$

where σ is a similarity function that maps image patch embeddings to values between 0 and 1. In the implementation, multi-head attention is employed in the matching module. We apply this matching process among all the tasks. Finally, the label decoder h projects the embeddings to the results.

Once trained, UniAP can easily adapt to *unseen* animal images at test time. This allows our model to adapt robustly to unseen animal images with a small prompt set. In brief, UniAP has a unified architecture and shares most of the parameters across tasks that can acquire generalizable priors, improving its few-shot learning capabilities.

Architecture

Our model follows a hierarchical encoder-decoder architecture that implements similarity matching with four components: image encoder $f_{\mathcal{T}}$, label encoder g, the matching module M, and label decoder h. Given the query image and the prompt set, the image encoder extracts patch-level embeddings (*tokens*) of each query and prompt image independently. The label encoder extracts tokens of each prompt label. Given the tokens at each hierarchy, the matching module performs matching to infer the tokens of the query label, from which the label decoder forms the raw query label.

Image Encoder Our image encoder relies on a Vision Transformer (ViT) (Dosovitskiy et al. 2020). The ViT processes each query and prompt image separately but with shared weights, resulting in a tokenized representation of

image patches at multiple levels. Following the approach in (Ranftl, Bochkovskiy, and Koltun 2021), we extract tokens from four intermediate ViT blocks to generate hierarchical features. To ensure that our system can learn general representations for a wide range of tasks, we initialize the parameters using pre-trained BEiT (Bao et al. 2021), which is self-supervised and less biased towards specific tasks.

For various tasks $\mathcal{T}$, we employ "bias tuning" (Cai et al. 2020; Zaken, Goldberg, and Ravfogel 2022). This involves the sharing of weights θ across all tasks, but the use of unique biases $\theta_{\mathcal{T}}$ for each individual task. By implementing this approach, we can effectively adjust to meet the demands of different tasks $\mathcal{T}$, ensuring optimal performance and accuracy.

Label Encoder The label encoder g is a lightweight function that utilizes only the patch embedding with a 2D convolution block and the positional embedding to extract tokens of each prompt label Y^q. Note that the label encoder g is shared across tasks.

Matching Module UniAP uses the feature pyramid network (FPN) (Lin et al. 2017) to maintain the affinity matrix through multi-head attention layers in each hierarchical matching module. We retrieve the tokens from the intermediate layers of image encoders for the query image X^q as $\{\mathbf{q}_j\}_{j\leq M}$ and prompt images X_i^p as $\{\mathbf{k}_{i\times j}\}_{i\leq N, j\leq M}$. Additionally, we obtain the tokens of prompt labels Y_i^p as $\{\mathbf{v}_{i\times j}\}_{i\leq N, j\leq M}$. Then, we organize the tokens into row vectors. Next, we utilize a multi-head attention layer to deduce the query label tokens at the hierarchy in the following manner:

$$\text{MHA}(\mathbf{q},\mathbf{k},\mathbf{v}) = \text{Concat}(\mathbf{o}_1,...,\mathbf{o}_h,...,\mathbf{o}_H)w^O, \quad (4)$$

where

$$\mathbf{o}_h = \text{Softmax}\left(\frac{\mathbf{q}w_h^Q(\mathbf{k}w_h^K)^\top}{\sqrt{d_H}}\right)\mathbf{v}w_h^V, \quad (5)$$

where H is number of heads, d_H is head size, and $w_h^Q, w_h^K, w_h^V \in \mathbb{R}^{d\times d_H}$, $w^O \in \mathbb{R}^{Hd_H\times d}$.

Note that each attention head in Eq. (5) plays the role of matching with lightweight functions. This is achieved by calculating the similarity between the query and prompt image tokens $\mathbf{q}$ and $\mathbf{k}$, which determines the weight given to each prompt label token $\mathbf{v}$ in inferring the query label token. The similarity function used is the scaled dot-product attention. The multi-head attention layer in Eq. (4) has several trainable projection matrices w_h^Q, w_h^K, w_h^V, which enables it to learn different branches (heads) of the matching algorithm, each with unique similarity functions.

Label Decoder The label decoder consists of a dense prediction head and binary classification head. In the dense prediction head, to predict the query label of the original resolution, the dense prediction head combines query label tokens inferred at multiple hierarchies. We choose the multi-scale decoder architecture of Dense Prediction Transformer (Ranftl, Bochkovskiy, and Koltun 2021) as it is compatible with ViT encoders and multi-level tokens. At each hierarchy of the decoder, the inferred query label tokens are first spatially concatenated to create a feature map of constant size. Then, (transposed) convolution layers of varying strides are used to produce a feature pyramid of increasing resolution for each feature map. The multi-scale features are then progressively upsampled and fused by convolutional blocks, followed by a convolutional head for final prediction. The dense prediction head shares all its parameters across tasks, enabling it to develop a versatile approach to decoding a structured label from predicted query label tokens. In addition, the dense prediction head output is single-channel, making it suitable for various tasks involving any number of channels. The binary classification head takes the output of the image encoder directly, followed by a linear feed-forward layer to obtain the classification score output.

Training and Inference

In the experiment, each dataset is split into two separate subsets by animal classes, namely $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, which does not have overlapping species. We train our model on a labeled dataset $\mathcal{D}_{\text{train}}$ of tasks $\mathcal{T} = \{PE, SS, CLS\}$. At each episode, we sample two labeled sets $\mathcal{P}_{\text{train}}, \mathcal{Q}_{\text{train}}$ from $\mathcal{D}_{\text{train}}$. Then we train the model to predict labels in $\mathcal{Q}_{\text{train}}$ using $\mathcal{P}_{\text{train}}$ as a prompt set. Throughout universal training, we expose the model to tasks like PE, SS, and CLS in the training data set $\mathcal{D}_{\text{train}}$. This allows the model to gain a comprehensive understanding of all the tasks and be capable of handling few-shot input data. Besides, we utilize a principled multi-task loss incorporating homoscedastic task uncertainty (Kendall, Gal, and Cipolla 2018) to learn multiple classification and regression losses with varying quantities and units.

Let $\mathcal{F}(X^q;\mathcal{P})$ denote the prediction of the model on X^q using the prompt set $\mathcal{P}$. We construct the homoscedastic task uncertainty through probabilistic modeling to determine the most efficient distribution of varying task weights w_i. Then the model $(f_{\mathcal{T}}, g, \mathcal{M}, h)$ is trained by the following learning objective in end-to-end:

$$\min_{f_{\mathcal{T}},g,\mathcal{M},h} \sum_{(X^q,Y^q)\in\mathcal{Q}_{\text{train}}} w_t\mathcal{L}_t\left(Y^q, \mathcal{F}(X^q;\mathcal{P}_{\text{train}})\right) + \log w_t, \quad (6)$$

Where the variable $\mathcal{L}$ represents the loss function, the parameter w_t represents the amount of noise in the task $t \in \mathcal{T}$. As the noise level increases, the weight of the corresponding loss function $\mathcal{L}_t$ decreases. Conversely, as the noise level decreases, the weight of the objective increases. To prevent the noise level from becoming too high and ignoring the data, the final term in the objective acts as a regulator for the noise parameters. In our experiments, we use the cross-entropy loss for all tasks. During training, we train separate sets of bias parameters for the image encoder $f_{\mathcal{T}}$ for each training task. In the pose estimation task, we ensure that each prompt label of a single image contains only one keypoint.

Once the model is trained on $\mathcal{D}_{\text{train}}$, we use it to evaluate the *unseen* animal species data $\mathcal{D}_{\text{test}}$ with a given prompt set $\mathcal{P}_{\text{test}}$. We adapt the model by fine-tuning the bias parameters of the image encoder $f_{\mathcal{T}}$ using the prompt set $\mathcal{P}_{\text{test}}$. To do this, we simulate episodic meta-learning by randomly dividing the prompt set into a sub-prompt set $\tilde{\mathcal{P}}$ and a sub-query

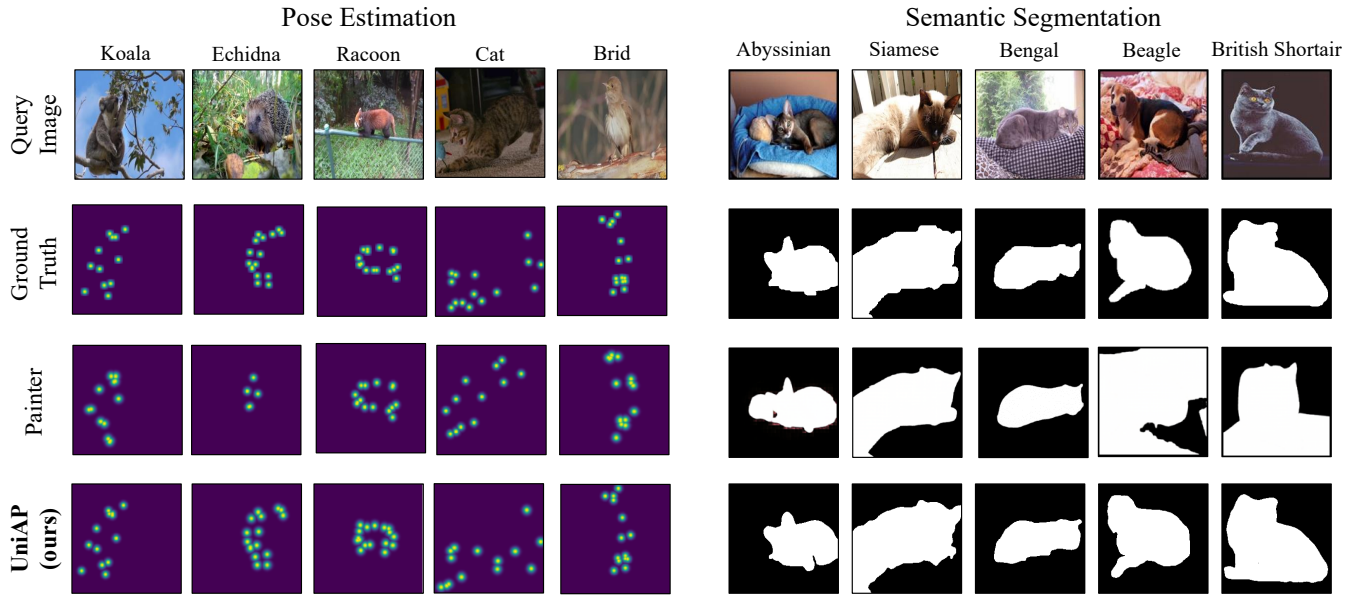

Figure 3: Qualitative Comparison. The comparison of representative methods in 1/0-shot evaluation for pose estimation in the Animal Kingdom and semantic segmentation in the Oxford-IIIT dataset.

set $\tilde{\mathcal{Q}}$. $\mathcal{P}_{\text{test}} = \tilde{\mathcal{P}} \dot{\cup} \tilde{\mathcal{Q}}$.

$$\min_{\theta_{\mathcal{T}}} \frac{1}{|\tilde{\mathcal{Q}}|} \sum_{(X^q, Y^q) \in \tilde{\mathcal{Q}}} \mathcal{L}\left(Y^q, \mathcal{F}(X^q; \tilde{\mathcal{P}})\right), \tag{7}$$

where $\theta_{\mathcal{T}}$ denotes bias parameters of the image encoder $f_{\mathcal{T}}$. As part of our implementation, we adjust the input projection of the label encoder g and the output head of the label decoder h, which has been shown to improve performance based on empirical evidence. These adjustments only require a small portion of parameters to be fine-tuned, which helps to prevent over-fitting on the limited prompt set $\mathcal{P}_{\text{test}}$. Once the model has been fine-tuned, we evaluate its effectiveness by using it to predict the label of an unseen query image with the prompt set $\mathcal{P}_{\text{test}}$.

Experiments

Experiment Setting

Implementation Details We conduct the experiment using $8\times$ NVIDIA RTX 3090. During the training process, we randomly pair image prompts and queries of the same animal category. We crop the objects of interest based on their bounding box and resize them to 224×224. To train our model, we utilize the Adam optimizer (Kingma and Ba 2015) for 1K iterations for warmup and 200K iterations in total. Our learning rate schedule follows the *poly* (Liu, Rabinovich, and Berg 2015) method with 0.9 decay rate, with base learning rates of 10^{-5} for pre-trained parameters and 10^{-4} for parameters trained from scratch. We stop the training based on the validation metric threshold. Our global batch size is 64, with different tasks randomly sampled for each batch. We also include prompt and query sets in each batch of a size of 5 for each. We adopt $\text{BEiT}_{\text{Base}}$ (Bao et al. 2021) backbone as the image encoder, which is pre-trained on ImageNet-22k (Deng et al. 2009).

Datasets We apply UniAP on three datasets for animal perception tasks: Animal Kingdom (Ng et al. 2022), Animal Pose (Cao et al. 2019), and APT-36K (Yang et al. 2022) for pose estimation. These datasets provide us with images of up to 49 kinds of animals with keypoint annotations. Animal Kingdom dataset is also used for the classification task. Additionally, we utilize the Oxford-IIIT Pet dataset (Parkhi et al. 2012) for the segmentation task. The Oxford-IIIT Pet is a 37-category dataset with segmentation annotations. Note that the categories of the train/test/val set are mutually exclusive. We split several animal classes from the raw datasets, 12 from Animal Kingdom, 2 from Animal Pose, 6 from APT-36K, 5 from Oxford-IIIT Pet.

Compared Baselines As no prior few-shot methods are developed for universal prediction tasks, we adapt state-of-the-art few-shot segmentation methods to our setup. We choose three models, Painter (Wang et al. 2023), POMNet (Xu et al. 2022), and HRNet-w48 (Wang et al. 2020). Painter is comprehensive on multi-tasks, while both POMNet and HRNet-w48 are specialized in pose estimation. In addition, we adopt Painter (Wang et al. 2023) and SAM (Kirillov et al. 2023) for prompt learning comparison, where Painter uses in-context prompting, while SAM needs user-interaction prompting. To validate the classification task, we compare our method (UniAP) with the clip-vit-large-patch14 (Radford et al. 2021) to perform zero-shot classification on the animal dataset.

Evaluation Protocol For pose estimation, we follow the PCK (Probability of Correct Keypoint) (Andriluka et al. 2014) . We report the average PCK ($\sigma = 0.05$ and 0.2) of

Model	# shots	Animal Kingdom		Animal Pose		APT-36K	
		PCK@0.2	PCK@0.05	PCK@0.2	PCK@0.05	PCK@0.2	PCK@0.05
$HRNet_{w48}$	-	90.49	62.04	**90.47**	75.91	91.65	66.26
Painter	1	70.52	48.34	77.86	53.85	74.11	51.39
POMNet	1	59.97	30.65	73.28	51.81	63.90	38.52
	3 / 3 / 2	79.15	52.88	77.7	49.96	5.79	38.4
UniAP (ours)	1	64.44	34.73	76.67	47.31	85.31	61.72
	30 / 35 / 40	**99.65**	**98.59**	90.10	**77.78**	**96.47**	**86.18**

Table 1: Pose estimation on Animal Kingdom, Animal Pose, and APT-36K datasets. We list the one-shot performance and the best performance with its number of shots on the bottom line for each dataset. Note that Painter can only use one-shot, while ours can use multi-shot to boost the performance.

Model	# shots	Acc.	mIoU
SAM_{user}	-	92.06	88.99
Painter	1	86.47	77.72
UniAP (ours)	1	97.08	93.38
	10	**97.11**	**94.27**

Table 2: Segmentation on Oxford-IIIT dataset. We list the one-shot performance and the best performance with its number of shots on the bottom line.

Model	# shots	Acc.
$CLIP_{ViT\text{-}Base}$	-	88.54
$CLIP_{ViT\text{-}Large}$	-	90.37
UniAP (ours)	1	92.19
	5	**93.13**

Table 3: Classification on Animal Kingdom dataset. We list the one-shot performance and the best performance with its number of shots on the bottom line.

all the categories in each split. For semantic segmentation, we adopt pixel accuracy and binary segmentation protocols and report both the standard accuracy (Acc.) and mean intersection over union (mIoU) for all classes. For classification, we utilize standard accuracy (Acc.).

Main Results

Pose Estimation As shown in Table 1, UniAP achieves state-of-the-art performance when the optimal number of shots is selected by few-shot on pose estimation, which reaches to or exceeds the performance of fully supervised models, far ahead of prompt learning or in-context learning methods such as POMNet (Xu et al. 2022) and Painter (Wang et al. 2023). Note that Painter can only use one-shot, while UniAP can further use multi-shot to boost the performance.

Segmentation As shown in Table 2, UniAP completely surpasses the similar prompt learning method Painter (Wang et al. 2023) and even the foundation model SAM (Kirillov et al. 2023) in semantic segmentation.

Classification As shown in Table 3, UniAP completely exceeds the performance of the Foundation model CLIP (Radford et al. 2021) with different parameter sizes.

Qualitative Comparison As shown in Fig. 3, UniAP demonstrates superior visual results compared to the Painter method, particularly in complex scenes where distinguishing between the target and background is challenging.

Ablation Study

In this section, we conduct extensive ablation studies to answer the following questions.

How does the number of shots matters? The results of our study can be seen in Fig. 4, where we demonstrate the effectiveness of UniAP. By gradually increasing the size of the prompt set from 1 to 40, our model reaches or outperforms both supervised methods and the foundation model across multiple tasks, even with a smaller amount of additional data. This suggests that UniAP could be particularly useful in specialized fields with limited labels.

Do we actually need the task-specific fine-tuning stage? Through our observations in Fig. 4, we have noticed that combining multi-task training and few-shot adaptation with an efficient parameter-sharing strategy for bias tuning significantly improves performance with a noticeable gap.

How does the domain affect performance? Based on the information provided in Table 4, we have observed three different settings: In Domain (ID), Out of Domain (OOD), and Cross Evaluation (CE). To be specific, ID refers to the prompt and query sets the same in the category. OOD and CE refer to different categories of prompt and query sets, where in CE prompt sets are sourced from the training set, while all of the prompt sets in OOD are from the test set and have not been trained. UniAP aims to maintain a consistent approach to animal perception tasks while also ensuring a comprehensive global understanding. This is achieved through task-specific fine-tuning. Additionally, we train on multiple large general-purpose animal datasets to ensure di-

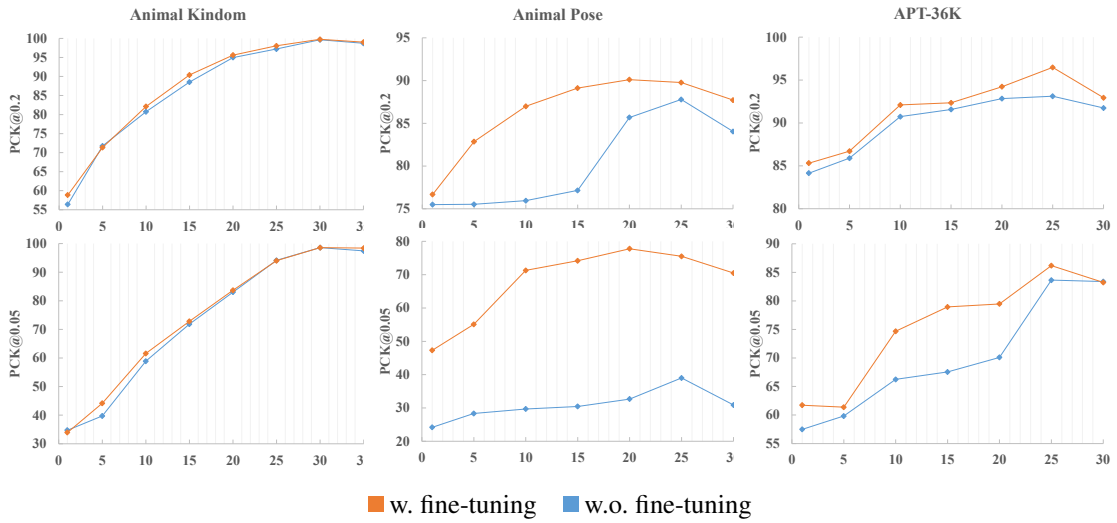

Figure 4: Ablation studies on pose estimation performance with models w/o finetuning. The x-axis indicates the number of shots. Note that the model performance is impacted by both the finetuning and the number of shots. The data quality is key to the initial performance and how it will perform over time. More shots mean receiving more prompt data, which improves its performance. The excessive number of shots will introduce additional noise and affect model performance. Thus, choosing the number of shots is a critical decision that can significantly impact the overall effectiveness of the model.

Setting	# shots	Animal Kingdom		Animal Pose		APT-36K	
		PCK@0.2	PCK@0.05	PCK@0.2	PCK@0.05	PCK@0.2	PCK@0.05
OOD	1	64.44	34.73	76.67	47.31	85.31	61.72
	30 / 35 / 40	**99.65**	**98.59**	90.10	77.78	96.47	86.18
ID	1	77.39	61.25	77.69	46.74	92.41	61.41
	20 / 20 / 35	99.26	98.10	**94.97**	**88.47**	**96.92**	**92.70**
CE	1	54.18	24.61	60.47	28.47	78.39	47.92
	5 / 5 / 5	65.50	28.86	77.33	50.62	88.72	71.67

Table 4: Ablation studies on evaluation settings. ID (in domain) refers to the prompt and query sets the same in the category. OOD (out of domain) and CE (cross evaluation) refer to different categories of prompt and query sets, where in CE prompt sets are sourced from the training set, while all of the prompt sets in OOD are from the test set and have not been trained. We list the one-shot performance and the best performance with its number of shots on the bottom line of each part.

versity. Our few-shot prompt mechanism allows for the efficient completion of tasks by utilizing prompt information.

Limitation and Future Work

Although UniAP employs a unified framework to handle animal pose estimation, segmentation, and classification tasks, several other tasks (such as object detection, behavior recognition, *etc.*) remain to be addressed in future work. Besides, exploring these tasks in a video-based context (*e.g.*, object tracking, video action recognition) is also worth investigating. Lastly, concerning animal perception, the quality and quantity of existing datasets are far from comparable to those of natural image datasets. The lack of fully annotated datasets hinders the sufficiency of our training. Therefore, collecting animal datasets or synthesizing through the virtual environment is a potential direction for future research. This would enable researchers to better understand the behavior of animals in their natural habitats.

Conclusion

This paper introduces UniAP, a universal animal perception vision model via few-shot learning, enabling the bridging of the gap between diverse animal species and visual tasks. Our proposed method integrating a transformer-based framework with task-specific bias tuning has demonstrated its efficacy in facilitating the cross-species animal perception learning process. Our proposed model takes support images and labels as prompt guidance for a query image. Images and labels are processed through a transformer-based encoder and a lightweight label encoder, respectively. Then the matching module aggregates information between prompt guidance and the query image, followed by a multi-head label decoder to generate outputs for various tasks.

Acknowledgements

This work is supported by the National Key R&D Program of China No.2022ZD0162000, and National Natural Science Foundation of China No.62106219.

References

Anderson, T. L.; and Donath, M. 1990. Animal behavior as a paradigm for developing robot autonomy. *Robotics and autonomous systems*, 6(1-2): 145–168.

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 3686–3693.

Ashwood, Z.; Jha, A.; and Pillow, J. W. 2022. Dynamic Inverse Reinforcement Learning for Characterizing Animal Behavior. *Advances in Neural Information Processing Systems*, 35: 29663–29676.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2021. BEiT: BERT Pre-Training of Image Transformers. In *ICLR*.

Butail, S.; Abaid, N.; Macrì, S.; and Porfiri, M. 2015. Fish–robot interactions: robot fish in animal behavioral studies. *Robot Fish: Bio-inspired Fishlike Underwater Robots*, 359–377.

Cai, H.; Gan, C.; Zhu, L.; and Han, S. 2020. Tinytl: Reduce memory, not parameters for efficient on-device learning. *NeurIPS*.

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

Dang, Q.; Yin, J.; Wang, B.; and Zheng, W. 2019. Deep learning based 2d human pose estimation: A survey. *Tsinghua Science and Technology*, 24(6): 663–676.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Fan, Q.; Pei, W.; Tai, Y.-W.; and Tang, C.-K. 2022. Self-support few-shot semantic segmentation. In *European Conference on Computer Vision*, 701–719. Springer.

Fan, Q.; Zhuo, W.; Tang, C.-K.; and Tai, Y.-W. 2020a. 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.

Fan, Z.; Yu, J.-G.; Liang, Z.; Ou, J.; Gao, C.; Xia, G.-S.; and Li, Y. 2020b. Fgn: Fully guided network for few-shot instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9172–9181.

Fang, Z.; and López, A. M. 2019. Intention recognition of pedestrians and cyclists by 2d pose estimation. *IEEE Transactions on Intelligent Transportation Systems*, 21(11): 4773–4783.

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.

Guo, Y.; Liu, Y.; Georgiou, T.; and Lew, M. S. 2018. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, 7: 87–93.

Iqbal, E.; Safarov, S.; and Bang, S. 2022. Msanet: Multi-similarity and attention guidance for boosting few-shot segmentation. *arXiv preprint arXiv:2206.09667*.

Jiang, L.; Lee, C.; Teotia, D.; and Ostadabbas, S. 2022. Animal pose estimation: A closer look at the state-of-the-art, existing gaps and opportunities. *Computer Vision and Image Understanding*, 103483.

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.

Kendall, A.; Gal, Y.; and Cipolla, R. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7482–7491.

Kim, D.; Kim, J.; Cho, S.; Luo, C.; and Hong, S. 2023. Universal Few-shot Learning of Dense Prediction Tasks with Visual Token Matching. In *International Conference on Learning Representations, ICLR 2023*. International Conference on Learning Representations, ICLR.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Lauer, J.; Zhou, M.; Ye, S.; Menegas, W.; Schneider, S.; Nath, T.; Rahman, M. M.; Di Santo, V.; Soberanes, D.; Feng, G.; et al. 2022. Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nature Methods*, 19(4): 496–504.

Li, C.; and Lee, G. H. 2021. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1482–1491.

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*, 17174–17183.

Li, L.; Dong, B.; Rigall, E.; Zhou, T.; Dong, J.; and Chen, G. 2021. Marine animal segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4): 2303–2314.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, W.; Rabinovich, A.; and Berg, A. C. 2015. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*.

López, J. G.; Agudo, A.; and Moreno-Noguer, F. 2019. Vehicle pose estimation via regression of semantic points of interest. In *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, 209–214. IEEE.

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.

Luo, H.; Bao, J.; Wu, Y.; He, X.; and Li, T. 2023. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, 23033–23044. PMLR.

Ma, C.; Yang, Y.; Wang, Y.; Zhang, Y.; and Xie, W. 2022. Open-vocabulary semantic segmentation with frozen vision-language models. *arXiv preprint arXiv:2210.15138*.

Michaelis, C.; Ustyuzhaninov, I.; Bethge, M.; and Ecker, A. S. 2018. One-shot instance segmentation. *arXiv preprint arXiv:1811.11507*.

Munea, T. L.; Jembre, Y. Z.; Weldegebriel, H. T.; Chen, L.; Huang, C.; and Yang, C. 2020. The progress of human pose estimation: A survey and taxonomy of models applied in 2D human pose estimation. *IEEE Access*, 8: 133330–133348.

Ng, X. L.; Ong, K. E.; Zheng, Q.; Ni, Y.; Yeo, S. Y.; and Liu, J. 2022. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19023–19034.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. V. 2012. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3498–3505.

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.

Pereira, T. D.; Tabris, N.; Matsliah, A.; Turner, D. M.; Li, J.; Ravindranath, S.; Papadoyannis, E. S.; Normand, E.; Deutsch, D. S.; Wang, Z. Y.; et al. 2022. SLEAP: A deep learning system for multi-animal pose tracking. *Nature methods*, 19(4): 486–495.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision transformers for dense prediction. In *ICCV*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Shaban, A.; Bansal, S.; Liu, Z.; Essa, I.; and Boots, B. 2017. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*.

Shi, X.; Yang, C.; Xia, X.; and Chai, X. 2020. Deep cross-species feature learning for animal face recognition via residual interspecies equivariant network. In *European Conference on Computer Vision*, 667–682. Springer.

Shooter, M.; Malleson, C.; and Hilton, A. 2021. SyDog: A synthetic dog dataset for improved 2D pose estimation. *arXiv preprint arXiv:2108.00249*.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3349–3364.

Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; and Cottrell, G. 2018. Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)*, 1451–1460. Ieee.

Wang, X.; Wang, W.; Cao, Y.; Shen, C.; and Huang, T. 2023. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6830–6839.

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, Y.; Yang, J.; Xu, Y.; Zhang, J.; Lan, L.; and Tao, D. 2022. Apt-36k: A large-scale benchmark for animal pose estimation and tracking. *Advances in Neural Information Processing Systems*, 35: 17301–17313.

Yu, H.; Xu, Y.; Zhang, J.; Zhao, W.; Guan, Z.; and Tao, D. 2021. AP-10K: A Benchmark for Animal Pose Estimation in the Wild. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Zaken, E. B.; Goldberg, Y.; and Ravfogel, S. 2022. BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Zhang, X.; Wang, W.; Chen, Z.; Xu, Y.; Zhang, J.; and Tao, D. 2023. CLAMP: Prompt-Based Contrastive Learning for Connecting Language and Animal Pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23272–23281.

Zhang, X.; Wang, W.; Chen, Z.; Zhang, J.; and Tao, D. 2022. Promptpose: Language prompt helps animal pose estimation. *arXiv preprint arXiv:2206.11752*.