CLAMP: Prompt-based Contrastive Learning for Connecting Language and Animal Pose

Xu Zhang\textsuperscript{1} Wen Wang\textsuperscript{2} Zhe Chen\textsuperscript{1} Yufei Xu\textsuperscript{1} Jing Zhang\textsuperscript{1} Dacheng Tao\textsuperscript{1}
\textsuperscript{1}The University of Sydney, Australia \textsuperscript{2}Zhejiang University, China
\{xzha0930, yuxu7116\}@uni.sydney.edu.au \ wwen@zju.edu.cn
\{zhe.chen1, jing.zhang1\}@sydney.edu.au \ dacheng.tao@gmail.com

Abstract

Animal pose estimation is challenging for existing image-based methods because of limited training data and large intra- and inter-species variances. Motivated by the progress of visual-language research, we propose that pre-trained language models (e.g., CLIP) can facilitate animal pose estimation by providing rich prior knowledge for describing animal keypoints in text. However, we found that building effective connections between pre-trained language models and visual animal keypoints is non-trivial since the gap between text-based descriptions and keypoint-based visual features about animal pose can be significant. To address this issue, we introduce a novel prompt-based Contrastive learning scheme for connecting Language and Animal Pose (CLAMP) effectively. The CLAMP attempts to bridge the gap by adapting the text prompts to the animal keypoints during network training. The adaptation is decomposed into spatial-aware and feature-aware processes, and two novel contrastive losses are devised correspondingly. In practice, the CLAMP enables the first cross-modal animal pose estimation paradigm. Experimental results show that our method achieves state-of-the-art performance under the supervised, few-shot, and zero-shot settings, outperforming image-based methods by a large margin. The code is available at https://github.com/xuzhang1199/CLAMP.

1. Introduction

Animal pose estimation aims to locate and identify a series of animal body keypoints from an input image. It plays a key role in animal behavior understanding, zoology, and wildlife conservation which can help study and protect animals better. Although the animal pose estimation task is analogous to human pose estimation \cite{2} to some extent, we argue that the two tasks are very different. For example, animal pose estimation involves multiple animal species, while human pose estimation only focuses on one category. Besides, it is much more difficult to collect and annotate animal pose data covering different animal species, thus existing animal pose datasets are several times smaller than the human pose datasets \cite{20} regarding the number of samples per species. Recently, Yu \textit{et al.} \cite{38} attempted to alleviate this problem by presenting the largest animal pose estimation dataset, i.e., AP-10K, which contains 10K images from 23 animal families and 54 species and provides the baseline performance of SimpleBaseline \cite{32} and HRNet \cite{31}. Despite this progress, the volume of this dataset is still far smaller than the popular human pose dataset, such as MS COCO \cite{20} with 200K images.

With diverse species and limited data, current animal pose datasets usually have large variances in animal poses which include both intra-species and inter-species vari-
ances. More specifically, the same animal can have diverse poses, e.g., pandas can have poses like standing, crawling, sitting, and lying down. Besides, the difference in the poses of different animal species can also be significant, e.g., horses usually lie down to the ground, while monkeys can be in various poses. Furthermore, even with the same pose, different animals would have different appearances. As an example, the joints of monkeys are wrinkled and hairy, while those of hippos are smooth and hairless. As a result, it could be extremely challenging for current human pose estimation methods to perform well on the animal pose estimation task without sufficient training data. Although image-based pre-training methodologies can be helpful in mitigating the problem of insufficient data, the huge gap between the pre-training datasets (e.g., ImageNet [7] image classification dataset and MS COCO human pose dataset [20]) and the animal pose datasets could compromise the benefits of pre-training procedures.

Rather than only using images to pre-train models, we notice that the keypoints of different poses and different animals share the same description in natural languages, thus the language-based pre-trained models can be beneficial to compensate for the shortage of animal image data. For example, if a pre-trained language model provides a text prompt of “a photo of the nose”, we can already use it to identify the presence of the nose keypoint in the image without involving too much training on the new dataset. Fortunately, a recently proposed Contrastive Language-Image Pre-training (CLIP) [28] model can provide a powerful mapping function to pair the image with texts effectively. Nevertheless, we found that fine-tuning the CLIP on the animal pose dataset could still suffer from large gaps between the language and the images depicting animals. In particular, the vanilla CLIP model only learns to provide a text prompt with general language to describe the entire image, while the animal pose estimation requires pose-specific descriptions to identify several different keypoints with their locations estimated from the same image. To this end, it is important to adapt the pre-trained language model to the animal pose dataset and effectively exploit the rich language knowledge for animal pose estimation.

To address the above issue, we propose a novel prompt-based contrastive learning scheme for effectively connecting language and animal pose (called CLAMP), enabling the first cross-modal animal pose estimation paradigm. In particular, we design pose-specific text prompts to describe different animal keypoints, which will be further embedded using the language model with rich prior knowledge. By adapting the pose-specific text prompts to visual animal keypoints, we can effectively utilize the knowledge from the pre-trained language model for the challenging animal pose estimation. However, there is a significant gap between the pre-trained CLIP model (which generally depicts the entire image) and the animal pose task (which requires the specific keypoint feature discriminative to and aligned with given text descriptions). To this end, we decompose the complicated adaptation into a spatial and feature-aware process. Specifically, we devise a spatial-level contrastive loss to help establish spatial connections between text prompts and the image features. A feature-level contrastive loss is also devised to make the visual features and embedded prompts of different keypoints more discriminative to each other and align their semantics in a compatible multi-modal embedding space. With the help of the decomposed adaptation, effective connections between the pre-trained language model and visual animal poses are established. Such connections with rich prior language knowledge can help deliver better animal pose prediction.

In summary, the contribution of this paper is threefold:

• We propose to decompose the cross-modal adaptation into a spatial-aware process and a feature-aware process with carefully designed losses, which could effectively align the language and visual features.
• We propose a novel cross-modal animal pose estimation paradigm named CLAMP to effectively exploit prior language knowledge from the pre-trained language model for better animal pose estimation.
• Experiments on two challenging datasets in three settings, i.e., 1) AP-10K [38] dataset (supervised learning, few-shot learning, and zero-shot learning) and 2) Animal-Pose [4] dataset (supervised learning), validate the effectiveness of the CLAMP method.

2. Related work

2.1. Pose estimation

Pose estimation is a challenging and active research area in computer vision. Most of the existing methods [5, 9, 12, 21, 25, 26, 31, 32, 35, 40, 40] focus on human pose estimation and simply predict the locations of keypoints based on images. Although they obtain superior performance for human pose estimation, these methods face difficulties in generalizing to animal pose estimation tasks, where there are large intra- and inter-species variances for different animal instances and limited training data per species. To tackle this problem, previous methods resort to domain adaptation or knowledge distillation [15, 36, 37] for animal pose estimation [4, 16, 24, 30, 38]. For example, Cao et al. [4] propose a cross-domain adaptation method, which transfers the knowledge in labeled human pose data to handle the unlabeled animal pose data. Li et al. [16] carry out the domain adaptation from synthetic to real data for animal pose estimation. Recently, Yu et al. [38] validate the effectiveness of leveraging pose estimation models pre-trained on human datasets for fine-tuning. However, they only focus on the
knowledge of image modality and still struggle to deal with multiple animal species with large variances in appearance, texture, and pose, especially in settings of limited data.

In this paper, we try to address this problem from a novel perceptive, i.e., using rich prior knowledge of language modality. We argue that although the features of keypoints from different animal images may have large variances, they share the same description in languages. Motivated by this, we propose a novel prompt-based contrastive learning scheme with decomposed spatial-aware and feature-aware adaptation processes for effectively connecting language and animal pose.

2.2. Vision-language models

Vision-language models cover a wide range of research topics [1,3,17,27,33], while we focus on reviewing the most related works on vision-language pre-training and fine-tuning. Vision-language pre-training has witnessed significant progress in the last few years, which generally learns an image encoder and a text encoder jointly [10,14,19,28]. A representative work is contrastive language–image pre-training dubbed CLIP [28], which uses 400 million text–image paired data to pre-train a multi-modal model. Experiments show that CLIP can help achieve effective few-shot or even zero-shot classification by simply exploring the relations between text features and image features.

Although significant progress has been made in vision-language pre-training, how to effectively adapt these pre-trained models to downstream tasks is still challenging and actively studied. For example, CoOp [42] and CoCoOp [43] take inspiration from prompt learning in NLP [22] and propose to utilize learnable text embedding for better image classification. Similarly, CLIP-adaptor [11] and TIP-adaptor [41] improve the model performance on downstream tasks through a lightweight adaptor. While the above methods focus on adapting CLIP for the image classification task, DenseCLIP [29] proposes a language-guided fine-tuning method for applying the pre-trained models to semantic segmentation and instance segmentation. GLIP [18] studies how to use image-text pairs to obtain a well pre-trained model that is suitable for object detection and phrase grounding. Different from them, our CLAMP makes the first attempt to leverage the language knowledge from the vision-language pre-trained model for animal pose estimation via specially designed pose-specific prompts and decomposed adaptation in both spatial and feature levels.

3. Method

3.1. Preliminary

Animal pose estimation pipeline Similar to the human pose estimation task, animal pose estimation aims to locate \( N \) keypoints of each animal instance in the input image. We follow most of the existing pose estimation methods and apply a typical top-down keypoint detection pipeline [31, 32], i.e., firstly using a detector to detect all animal instances in the image, then detecting the keypoints for each instance. Specifically, the heatmap representation is usually used to denote the location of each keypoint. We denote the cropped instance image as \( I \in \mathbb{R}^{h \times w \times 3} \), where \( h \) and \( w \) are the height and width of the image, respectively. The image encoder \( f_{\text{extr}} \) extracts the image feature \( F \in \mathbb{R}^{h_0 \times w_0 \times C} \) from \( I \), where \( h_0, w_0 \), and \( C \) are the height, width, and the number of channels of the extracted feature, respectively. An ImageNet [7] pre-trained backbone network, e.g., ResNet-50 [13] or HRNet-32 [31], is usually employed as the image encoder in image-based methods, while we adopt CLIP pre-trained ResNet-50 and ViT [8] in our CLAMP to leverage the language knowledge. A typical ratio \( s_0 \) between \( h \) and \( h_0 \) is 32 for ResNet-50 and 4 for HRNet-32. Then, the keypoint predictor \( f_{\text{pred}} \) decodes \( F \) into a heatmap \( H \in \mathbb{R}^{h_1 \times w_1 \times N} \), which typically consists of several deconvolution layers depending on the ratio \( s_0 \) and a convolutional prediction layer. The ratio \( s_1 \) between \( h \) and \( h_1 \) is 4. Finally, we get the coordinates of \( N \) keypoints \( K \in \mathbb{R}^{N \times 3} \) by applying a simple argmax operation on each heatmap and multiplying the coordinates with the scale ratio \( s_1 \) to recover to the original scale:

\[
K_n = s_1 \cdot \argmax_{1 \leq i \leq h_1, 1 \leq j \leq w_1} H_n(i, j), \quad n = 1, \ldots, N, \tag{1}
\]

where \( K_n \) is the 2D coordinate of \( n \)-th keypoint and \( H_n \) is \( n \)-th heatmap in \( H \).

Language model CLIP [28] provides visual-feature-compatible pre-trained language models by leveraging a large number of paired images and text descriptions in pre-training. By using natural language to reference learned visual concepts, CLIP enhances the generality of the pre-trained models, enabling effective knowledge transfer to downstream tasks. For example in classification, it uses “A photo of a/an {object}” as a template to formulate text prompts, where {object} can be filled with names of different categories. By calculating the similarity between image features and different embedded text prompts, the CLIP pre-trained models can directly adapt to the classification task because it is similar to the pre-training process. CLIP aligns visual concepts with languages and demonstrates effective knowledge adaptation to downstream tasks. Motivated by this, we propose to exploit the prior language knowledge in CLIP pre-trained models and design a novel cross-modal animal pose estimation paradigm named CLAMP.

3.2. CLAMP

Our CLAMP includes the introduction of a set of pose-specific text prompts and two decomposed adaptation processes for leveraging prior language knowledge. Fig. 2 illustrates the proposed CLAMP method.
3.2.1 Pose-specific text prompts

Different from the classification tasks explored in CLIP, pose estimation is agnostic to the category of the animal instance and needs to find the position of a set of local keypoints in each image. Thus, we need to design pose-specific text prompts for animal pose estimation. Motivated by CoOp [42], we use a learnable template of $k$ learnable prefix tokens instead of the fixed “A photo of a” template in CLIP to learn a prompt template that adapts better to pose estimation. Accordingly, we fill in \{object\} with the names of different keypoints like ‘nose’ to get $N$ text prompts, i.e.,

$$p_n = [T]_1[T]_2 ... [T]_n[\text{KeyPoint}]_n, \quad n = 1, ..., N,$$

where $[T]_i, i \in \{1, 2, ..., k\}$ represents the learnable prefix tokens, $[\text{KeyPoint}]_n$ represents the $n$-th keypoint name, and $N$ is the number of keypoints. The $N$ text prompts are mapped to the multi-modal embedding space by using a CLIP pre-trained text encoder to get the prompt embedding $E_{\text{prompt}}^{\text{origin}} \in \mathbb{R}^{N \times C_{\text{emb}}}$. Considering the intrinsic relationship between different keypoints is important for pose estimation, we further employ a lightweight prompt encoder (i.e., a single transformer layer) to model the relationship between the prompt embeddings of different keypoints and promote their interactions. After that, cross attention is applied to enhance the prompt embeddings with the image feature [29], generating the enhanced prompt embedding $E_{\text{prompt}} \in \mathbb{R}^{N \times C_{\text{emb}}}$. Although CLIP exhibits effective knowledge transfer ability in downstream classification with the help of its designed prompts, it is still challenging to directly adapt the text prompts to animal pose estimation due to the lack of spatial connection between text description and image feature in CLIP pre-training. To address this challenge, we decompose the cross-modal adaptation into a spatial-aware process and a feature-aware process. Furthermore, a spatial-level contrastive loss and a feature-level contrastive loss are devised to constrain the two processes, respectively.

Spatial-aware adaptation Spatial-aware adaptation aims at establishing spatial connections between the text prompts and image features, which can provide positional information for the animal pose. In this process, we devise a spatial-level contrastive loss to query the possibility of the presence of different animal keypoints in spatial dimension. Specifically, we feed the input image into the image encoder to obtain the image feature $F_{\text{origin}} \in \mathbb{R}^{H \times W \times C}$, where $H$, $W$, and $C$ represent the height, width, and the number of channels, respectively. Then the obtained $F_{\text{origin}}$ is mapped to the multi-modal embedding space through a projector, obtaining $F \in \mathbb{R}^{H \times W \times C_{\text{emb}}}$. For different image encoders, we employ slightly different projectors. For example, if the encoder takes the form of a ViT [8], the projector is a linear projection layer with the cls token as input. For the ResNet [13] encoder, the projector contains a global average pooling layer, a multi-head self-attention (MHSA) layer, and a linear projection layer, following the design in CLIP [38]. After that, the extracted prompt embeddings and image features are normalized and used to query the presence score map can help deliver effective animal pose estimation.

3.2.2 Pose-aware language knowledge adaptation

Figure 2. Conceptualized illustration of the proposed CLAMP method. The pipeline contains an image encoder extracting the visual feature of the input image, a text encoder encoding the text prompt with rich prior knowledge, and two adaptation modules for adapting prior language knowledge to visual animal pose. The obtained presence score map can help deliver effective animal pose estimation.
ence possibility of different keypoints via the inner product. Thus, we can get the presence score at each spatial position:

\[ S_{ijn} = F_{ij} \cdot E_{prompt}^n, \]
\[ i = 1, \ldots, H; j = 1, \ldots, W; n = 1, \ldots, N, \]

where \( F_{ij} \in \mathbb{R}^{1 \times C_{emb}} \) is the feature vector of \( F \) at pixel \((i, j)\), and \( E_{prompt}^n \) is the \( n \)-th prompt embedding in \( E_{prompt} \). Accordingly, we can get the presence score map by collecting and stacking the scores at all locations and prompts, i.e.,

\[ S = \text{Stack}(S_{ijn}), \quad S \in \mathbb{R}^{H \times W \times N}, \]

where \( \text{Stack} \) represents the stack operation.

Considering the effectiveness of Gaussian heatmap in pose estimation [31, 32], we use the target 2D Gaussian heatmap \( H_{\text{target}} \) to supervise the estimated score map via the following spatial-level contrastive loss, i.e.,

\[ L_{\text{spatial}} = \text{MSE}(\text{Upsample}(S), H_{\text{target}}), \]

where \( \text{MSE} \) is the mean squared error loss. We upsmaple the score map to align the spatial size of the score map and the target heatmap.

**Feature-aware adaptation** Given CLIP pre-training only learns to reference the entire image while animal pose requires more discriminative keypoint features to align with the corresponding text prompts, we introduce a feature-level contrastive loss for feature-aware adaptation. We encourage the visual feature of a specific keypoint to be close to the text prompt describing the corresponding keypoint and to be far away from those describing other keypoints, and apply the same operation on embedded text prompts, thereby enhancing the discriminative ability of the extracted text and image features and facilitating their alignment. Specifically, during the training process, we use the ground truth locations of the keypoint \( K \in \mathbb{R}^{N \times 2} \) to perform grid sampling on \( F \) to obtain the local visual features of \( N \) keypoints, i.e.,

\[ F_n \in \mathbb{R}^{1 \times C_{emb}}, \quad n = 1, \ldots, N. \]

The stacked keypoint feature can be obtained, i.e.,

\[ F_{\text{keypoint}} = \text{Stack}(F_n), \quad F_{\text{keypoint}} \in \mathbb{R}^{N \times C_{emb}}, \]

where \( \text{Stack} \) represents the stack operation. Then, we calculate the semantic matching score map between the visual feature of keypoints and prompt embeddings as follows:

\[ M = \hat{F}_{\text{keypoint}} \cdot E_{\text{prompt}}^T, \quad M \in \mathbb{R}^{N \times N}, \]

where \( \hat{F}_{\text{keypoint}} \) and \( E_{\text{prompt}} \) are normalized keypoint features and prompt embeddings, respectively. Since there is only one prompt embedding describing one given visual keypoint feature, we can simply use the diagonal matrix as the matching target \( M_{\text{label}} \). We perform contrastive learning on both prompt embeddings and keypoint features based on the following feature-level contrastive loss, i.e.,

\[ L_{\text{feature}} = \frac{1}{2}(CE(M, M_{\text{label}}) + CE(M^T, M_{\text{label}})), \]

where \( CE \) represents the cross entropy loss.

### 3.2.3 Final Prediction and Learning Objective

With the designed pose-specific text prompts and the decomposed cross-modal adaptation, our proposed CLAMP could connect text descriptions to visual features, making it possible to adapt the rich prior language knowledge from pre-trained language models to animal pose estimation. To let language knowledge collaborate with image features for animal pose estimation, we fuse the image features and the spatial presence score maps, i.e.,

\[ F_{\text{fuse}} = F_{\text{origin}} \oplus S, \quad F_{\text{fuse}} \in \mathbb{R}^{H \times W \times (C + N)}, \]

where \( \oplus \) represents the concatenate operation along the channel dimension. Then, \( F_{\text{fuse}} \) is fed into a keypoint predictor to predict the pose heatmap. The prediction results are supervised by a prediction loss \( L_{\text{pred}} \), which adopts the form of \( \text{MSE} \) loss between the predicted heatmap and ground truth heatmap. The overall training loss can be written as:

\[ L_{\text{total}} = L_{\text{pred}} + \alpha_1 \cdot L_{\text{spatial}} + \alpha_2 \cdot L_{\text{feature}}, \]

where \( \alpha_1 \) and \( \alpha_2 \) are two hyper-parameters to balance the importance of \( L_{\text{spatial}} \) and \( L_{\text{feature}} \).

### 4. Experiments

#### 4.1. Experimental setup

**Datasets and evaluation metrics** We employ the AP-10K [38] and Animal-Pose [4] datasets to evaluate the performance of the proposed CLAMP method. The AP-10K dataset contains 10,015 images collected and filtered from 23 animal families and 54 species, which is the largest and most diverse dataset for animal pose estimation. 17 keypoints are annotated in the dataset, i.e., two eyes, one nose, one neck, two shoulders, two elbows, two knees, two hips, four paws, and one tail. We adopt the training set during the training process and evaluate the model’s performance on the validation set. On the other hand, the Animal-Pose dataset covers 5 different animal species with over 4,000 images. 20 keypoints are annotated in each animal instance, including two eyes, throat, nose, withers, two earbases, one base of the tail, four elbows, four knees, and four paws. Similarly, we adopt the training set for training and report the results on the validation set. Following the common practice in animal pose estimation, we adopt the average precision (AP) as the main metric on the two datasets, which is computed based on the object keypoint similarity (OKS). The detailed protocol definitions can be found in [32].

**Implementation details** We employ the widely used two-stage top-down pose estimation paradigm similar to SimpleBaseline [32] in the experiments. The ground truth
<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Pre-train</th>
<th>AP</th>
<th>AP50</th>
<th>AP75</th>
<th>APM</th>
<th>APL</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimpleBaseline</td>
<td>ResNet-50</td>
<td>ImageNet</td>
<td>70.2</td>
<td>94.2</td>
<td>76.0</td>
<td>45.5</td>
<td>70.4</td>
<td>73.5</td>
</tr>
<tr>
<td>SimpleBaseline</td>
<td>ResNet-50</td>
<td>CLIP</td>
<td>70.9</td>
<td>94.6</td>
<td>76.8</td>
<td>44.8</td>
<td>71.2</td>
<td>74.1</td>
</tr>
<tr>
<td>CLAMP (ours)</td>
<td>ResNet-50</td>
<td>CLIP</td>
<td>72.9</td>
<td>95.4</td>
<td>79.4</td>
<td>43.2</td>
<td>73.2</td>
<td>76.3</td>
</tr>
<tr>
<td>SimpleBaseline</td>
<td>ViT-Base</td>
<td>CLIP</td>
<td>72.6</td>
<td>94.7</td>
<td>79.5</td>
<td>43.4</td>
<td>72.8</td>
<td>75.8</td>
</tr>
<tr>
<td>CLAMP (ours)</td>
<td>ViT-Base</td>
<td>CLIP</td>
<td>74.3</td>
<td>95.8</td>
<td>81.4</td>
<td>47.6</td>
<td>74.9</td>
<td>77.5</td>
</tr>
</tbody>
</table>

Table 1. Performance comparison on AP-10K [38].

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Pre-train</th>
<th>AP</th>
<th>AP50</th>
<th>AP75</th>
<th>APM</th>
<th>APL</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimpleBaseline</td>
<td>ResNet-50</td>
<td>ImageNet</td>
<td>51.1</td>
<td>85.4</td>
<td>50.2</td>
<td>26.7</td>
<td>51.3</td>
<td>56.3</td>
</tr>
<tr>
<td>SimpleBaseline</td>
<td>ResNet-50</td>
<td>CLIP</td>
<td>51.8</td>
<td>84.3</td>
<td>52.5</td>
<td>26.1</td>
<td>52.0</td>
<td>56.9</td>
</tr>
<tr>
<td>CLAMP (ours)</td>
<td>ResNet-50</td>
<td>CLIP</td>
<td>54.0</td>
<td>85.9</td>
<td>56.2</td>
<td>26.4</td>
<td>54.2</td>
<td>58.9</td>
</tr>
<tr>
<td>SimpleBaseline</td>
<td>ViT-Base</td>
<td>CLIP</td>
<td>57.4</td>
<td>88.9</td>
<td>61.2</td>
<td>20.9</td>
<td>57.8</td>
<td>61.2</td>
</tr>
<tr>
<td>CLAMP (ours)</td>
<td>ViT-Base</td>
<td>CLIP</td>
<td>61.2</td>
<td>91.3</td>
<td>64.7</td>
<td>36.8</td>
<td>61.7</td>
<td>65.2</td>
</tr>
</tbody>
</table>

Table 2. 20-shot performance comparison on AP-10K [38].

Bounding box annotations are utilized in AP-10K to crop animal instances during the training and evaluation process, following the default setting in [38]. We select the widely used ResNet [13] and the recently introduced attention-based vision transformer ViT [8] as the backbone networks for image feature extraction, i.e., ResNet-50 and ViT-Base. We use the ImageNet [7] pre-trained and CLIP [28] pre-trained weights to initialize the backbone networks for evaluating the effect of different pre-training methods. The text encoders after CLIP pre-training, i.e., CLIP-ResNet-50 and CLIP-ViT-Base, are adopted as our language models and initialized with the corresponding CLIP pre-trained weights. The decoder described in SimpleBaseline [32] is employed to predict the keypoints, which contains several deconvolution layers to upsample the extracted features to 1/4 of the input resolution and one convolution layer with kernel size 1 × 1 to predict the heatmap.

During training, we follow most of the training settings in AP-10K, i.e., the input image of each instance was cropped and resized to 256 × 256, followed by random flip, rotation, and scale jitter. Each model is trained for a total of 210 epochs with a step-wise learning rate schedule which decays by 10 at the 170th and 200th epoch, respectively. We use the AdamW [23] optimizer with a weight decay of 1e-4. Furthermore, we conduct supervised learning, few-shot learning, and zero-shot learning to thoroughly evaluate the models’ performance. For supervised learning on AP-10K and Animal-Pose, we train the model with a batch size of 128 and set 5e-4 as the initial learning rate. An extra learning rate multiplier of 0.1 is applied to the backbone weights to prevent over-fitting when using the ViT model as the backbone. The few-shot and zero-shot learning are conducted on AP-10K as it has much more animal species than Animal-Pose. For few-shot learning, we adopt a batch size of 64 and set the initial learning rate as 5e-4/5e-5 for ResNet-50/ViT-Base. For zero-shot learning configurations, we use a batch size of 128 and set 5e-4 as the initial learning rate. In all experiments, we set k to 8 in Eq. 2 and freeze the text encoder to reduce the computational cost. We show complexity analysis in the supplementary material.

4.2. Results and analysis

4.2.1 Experiments on AP-10K

Supervised learning The results under the supervised learning setting on AP-10K are shown in Table 1. It can be observed that the CLIP pre-training can help deliver better results on the animal pose estimation task than using the ImageNet pre-training, e.g., SimpleBaseline [32] with the CLIP pre-trained ResNet-50 backbone obtains 0.7 AP higher than the counterpart with the ImageNet pre-training. It validates that the image model trained with the prior language knowledge can benefit in dealing with the inter- and intra-species variance in animal pose estimation and thus bringing performance improvements. Furthermore, with the help of the proposed CLAMP, the model achieves much better performance, i.e., there is a performance gain of 2 AP than directly using the CLIP pre-trained model. Such observation demonstrates that with the proposed pose-specific prompts and decomposed adaptation, the language knowledge is better exploited in the animal pose estimation task and brings better performance. Similar conclusion can be drawn by observing the results using the ViT-Base backbone, e.g., the proposed CLAMP outperforms SimpleBaseline by 1.7 AP. Compared with the representative methods [26,31,32] that report results on AP-10K [38] and recent pose estimation methods [34,39], i.e., the results in Table 5, the proposed CLAMP model using ViT-Large as backbone achieves state-of-the-art performance, showing the potential of the proposed CLAMP method, especially given the good scalability of the model size of ViT.

Few-shot learning We also conduct few-shot learning experiments to study the generalization ability of different methods. We randomly selected 20 samples from each species in the training set of AP-10K to form a 20-shot animal pose estimation training set. The model is tested
<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Train</th>
<th>Test</th>
<th>(AP)</th>
<th>(AP_{50})</th>
<th>(AP_{75})</th>
<th>(AP_M)</th>
<th>(AP_L)</th>
<th>(AR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimpleBaseline [32]</td>
<td>ResNet-50</td>
<td>Bovidae</td>
<td>Canidae</td>
<td>41.3</td>
<td>79.4</td>
<td>36.4</td>
<td>26.8</td>
<td>41.3</td>
<td>49.1</td>
</tr>
<tr>
<td>CLAMP (ours)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SimpleBaseline [32]</td>
<td>ResNet-50</td>
<td>Canidae</td>
<td>Felidae</td>
<td>46.9</td>
<td>84.4</td>
<td>45.6</td>
<td>30.3</td>
<td>46.9</td>
<td>53.8</td>
</tr>
<tr>
<td>CLAMP (ours)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Zero-shot performance comparison on AP-10K [38].

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Pre-train</th>
<th>Test</th>
<th>(AP)</th>
<th>(AP_{50})</th>
<th>(AP_{75})</th>
<th>(AP_M)</th>
<th>(AP_L)</th>
<th>(AR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimpleBaseline [32]</td>
<td>ResNet-50</td>
<td>ImageNet</td>
<td></td>
<td>68.7</td>
<td>93.7</td>
<td>76.9</td>
<td>63.7</td>
<td>69.9</td>
<td>73.0</td>
</tr>
<tr>
<td>SimpleBaseline [32]</td>
<td>ResNet-50</td>
<td>CLIP</td>
<td></td>
<td>70.8</td>
<td>94.8</td>
<td>79.5</td>
<td>67.3</td>
<td>72.0</td>
<td>75.0</td>
</tr>
<tr>
<td>CLAMP (ours)</td>
<td>ResNet-50</td>
<td>CLIP</td>
<td></td>
<td>72.5</td>
<td>94.8</td>
<td>81.7</td>
<td>67.9</td>
<td>73.8</td>
<td>76.7</td>
</tr>
<tr>
<td>SimpleBaseline [32]</td>
<td>ViT-Base</td>
<td>CLIP</td>
<td></td>
<td>72.3</td>
<td>94.7</td>
<td>82.1</td>
<td>69.4</td>
<td>73.3</td>
<td>76.5</td>
</tr>
<tr>
<td>CLAMP (ours)</td>
<td>ViT-Base</td>
<td>CLIP</td>
<td></td>
<td>74.3</td>
<td>95.8</td>
<td>83.4</td>
<td>71.9</td>
<td>75.2</td>
<td>78.3</td>
</tr>
</tbody>
</table>

Table 4. Performance comparison on Animal-Pose [4].

<table>
<thead>
<tr>
<th>Method</th>
<th>(AP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimpleBaseline [32]</td>
<td>69.9</td>
</tr>
<tr>
<td>Hourglass [26]</td>
<td>72.9</td>
</tr>
<tr>
<td>HRNet-w32 [31]</td>
<td>73.8</td>
</tr>
<tr>
<td>HRNet-w48 [31]</td>
<td>74.4</td>
</tr>
<tr>
<td>ViPNAS* [34]</td>
<td>67.1</td>
</tr>
<tr>
<td>HRFormer-S* [39]</td>
<td>71.7</td>
</tr>
<tr>
<td>HRFormer-B* [39]</td>
<td>73.5</td>
</tr>
<tr>
<td>CLAMP-ResNet-50 (ours)</td>
<td>72.9</td>
</tr>
<tr>
<td>CLAMP-ViT-Base (ours)</td>
<td>74.3</td>
</tr>
<tr>
<td>CLAMP-ViT-Large (ours)</td>
<td>77.8</td>
</tr>
</tbody>
</table>

Table 5. Comparison with previous methods in AP-10K [38]. * indicates the results using the official mmpose [6] implementation.

4.2.2 Experiments on Animal-Pose

In addition to AP-10K, we further evaluate the effectiveness of CLAMP on Animal-Pose [4] dataset, which has a different data distribution compared with AP-10K. To train the CLAMP model using the data and annotations from the Animal-Pose dataset, we replace \([KeyPoint]\) in Eq. (2) with the keypoint names defined in Animal-Pose and expand the number of different keypoints to 20. The training setting is the same as the supervised setting described in the previous section. As shown in Table 4, the benefit of using CLIP pre-training is consistent in both the AP-10K dataset and Animal-Pose dataset, e.g., there is an improvement of 5.6 AP and 8.8 AP with CLAMP in these two settings, respectively. Such observation well demonstrates that language knowledge can greatly improve the models’ generalization ability since the shared language knowledge of keypoints can alleviate the difficulties caused by large visual inter- and intra-species variances.
trained under the supervised learning setting for 210 epochs. The variants are contrastive loss, the feature-level contrastive loss, and the CLAMP method in this section, i.e., the spatial-level adaptation, each keypoint feature has the highest similarity with the corresponding text prompt (i.e., the diagonal elements). It demonstrates that the feature-level adaptation helps enhance the discrimination of prompt embeddings and visual features of different keypoints, leading to better cross-modal alignment.

**Spatial-level score map and Qualitative analysis** We visualize the score map to study the effect of the spatial-level contrastive loss in the supplementary material, which shows the established connections between language descriptions and image features and can help understand our CLAMP. In addition, some visual results are presented in the supplementary material, showing the superiority of our method over the baseline model.

### 4.3. Ablation study

We use ResNet-50 [13] as the backbone network for pose estimation and ablate the effectiveness of the key designs in the CLAMP method in this section, i.e., the spatial-level contrastive loss, the feature-level contrastive loss, and the prompt encoder in pose-specific prompts. The variants are trained under the supervised learning setting for 210 epochs.

<table>
<thead>
<tr>
<th>$L_{\text{spatial}}$</th>
<th>$L_{\text{feature}}$</th>
<th>PromptEncoder</th>
<th>AP</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>×</td>
<td>×</td>
<td>×</td>
<td>70.9</td>
<td>74.1</td>
</tr>
<tr>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>72.1</td>
<td>75.3</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>72.6</td>
<td>75.8</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>72.9</td>
<td>76.3</td>
</tr>
</tbody>
</table>

Table 6. Ablation study of CLAMP with a CLIP pre-trained ResNet-50 backbone on AP-10K [38].

Figure 3. Visualization of the feature-level score map. For each grid-sampled keypoint feature (vertical axis), we calculate its similarity with different text prompts (horizontal axis) after CLIP pre-training (left) and after fine-tuning with the feature-level loss (right). Note that the left eye (i.e., the first keypoint) of the animal is invisible in the test image.

**4.4. Visualization and analysis**

*Feature-level score map* Based on Eq. (7), we can obtain the similarities between text prompts and local keypoint features before and after training with the feature-level loss $L_{\text{feature}}$. As shown in Fig. 3, the keypoint feature without feature-level adaptation has almost the same similarity for different text prompts, demonstrating the lack of discrimination in the visual keypoint features that are directly extracted by CLIP pre-trained models. With the feature-level adaptation, each keypoint feature has the highest similarity with the corresponding text prompt (i.e., the diagonal elements). It demonstrates that the feature-level adaptation helps enhance the discrimination of prompt embeddings and visual features of different keypoints, leading to better cross-modal alignment.

**Spatial-level score map and Qualitative analysis** We visualize the score map to study the effect of the spatial-level contrastive loss in the supplementary material, which shows the established connections between language descriptions and image features and can help understand our CLAMP. In addition, some visual results are presented in the supplementary material, showing the superiority of our method over the baseline model.

### 5. Conclusion and discussion

This paper proposes CLAMP to introduce prior language knowledge into animal pose estimation. With pose-specific prompts and the spatial-aware and feature-aware adaptation processes, CLAMP provides a promising solution to the well-known challenge in animal pose estimation, i.e., large intra- and inter-species variances together with limited data per species. Extensive experiments on the AP-10K and Animal-Pose benchmarks demonstrate that CLAMP outperforms representative methods by a large margin in all the supervised, few-shot, or zero-shot learning settings. As the first study of cross-modal animal pose estimation, we hope it can provide valuable insights and draw attention from the research community to improve animal pose estimation by effectively exploiting multi-modal knowledge. Besides, the proposed method can also benefit human pose estimation, especially in low-data regimes, which is presented in the supplementary material.

**Limitation discussion** In this study, we only adopt language models pre-trained on the CLIP dataset, which contains language-image pairs for various scenarios. In the future, we plan to investigate the influence of using an animal pose-related text-image pair dataset for multi-modal pre-training as well as develop more effective visualization tools to explain the learning process and the predictions.

**Acknowledgement** Mr Xu Zhang, Dr Zhe Chen, Mr Yufei Xu, and Dr Jing Zhang were supported by Australian Research Council Projects in part by FL170100117 and IH180100002.
References


