OpenKD: Opening Prompt Diversity for Zero- and Few-shot Keypoint Detection

Changsheng Lu[1], Zheyuan Liu[1], and Piotr Koniusz[⋆,2,1]

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

Abstract. Exploiting foundation models (*e.g.*, CLIP) to build a versatile keypoint detector has gained increasing attention. Most existing models accept either the *text prompt* (*e.g.*, "the nose of a cat"), or the *visual prompt* (*e.g.*, support image with keypoint annotations), to detect the corresponding keypoints in query image, thereby, exhibiting either *zero-shot* or *few-shot* detection ability. However, the research on multimodal prompting is still underexplored, and the prompt diversity in semantics and language is far from opened. For example, how to handle unseen text prompts for novel keypoint detection and the diverse text prompts like "Can you detect the nose and ears of a cat?" In this work, we open the prompt diversity in three aspects: *modality*, *semantics* (seen *vs.* unseen), and *language*, to enable a more general zero- and few-shot keypoint detection (Z-FSKD). We propose a novel OpenKD model which leverages a multimodal prototype set to support both visual and textual prompting. Further, to infer the keypoint location of unseen texts, we add the auxiliary keypoints and texts interpolated in visual and textual domains into training, which improves the spatial reasoning of our model and significantly enhances zero-shot novel keypoint detection. We also find large language model (LLM) is a good parser, which achieves over 96% accuracy when parsing keypoints from texts. With LLM, OpenKD can handle diverse text prompts. Experimental results show that our method achieves state-of-the-art performance on Z-FSKD and initiates new ways of dealing with unseen text and diverse texts. The source code and data are available at `https://github.com/AlanLuSun/OpenKD`.

1 Introduction

Keypoint detection is a fundamental research problem in computer vision and has numerous applications such as pose estimation for humans [6, 8, 12, 33, 37, 52] and animals [2, 5, 23, 34], action recognition [30, 47], and fine-grained image classification [39, 55]. Over the past decade, significant advancements have been made in deep keypoint detection. However, most existing methods are tailored to close-set detection and struggle to predict novel body parts and species with limited data. Thus, it necessitates such a zero- or few-shot keypoint detection:

⋆ Corresponding author.

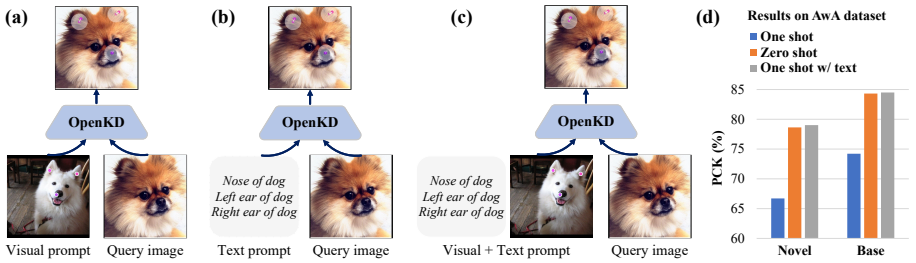


Fig. 1: Illustration of multimodal prompting for keypoint detection. Our model can successfully detect keypoints given visual prompts formed by support images and keypoints (a), text prompts (b), or both (c). Graph (d) shows our model well combines the advantages of different modalities, mitigating the weakness induced by either modality.

after being trained on a diverse dataset, the model can quickly recognize novel or base keypoints in unseen species given only *zero, one or a few labelled samples*.

On the one hand, researchers have recently explored few-shot keypoint detection (FSKD) [27–29, 53]. The FSKD inspired by few-shot learning [13, 20, 36, 38, 44] is general and offers higher flexibility of detecting a varying number of keypoints in a query image, given the prompts formed by support image with keypoint annotations. In this regard, the support set is namely visual prompt, which is required when evaluating new classes. On the other hand, in the era of foundation model, the vision-language model (*e.g.*, CLIP [35]) injects new life to detecting keypoints, enabling zero-shot keypoint detection (ZSKD) [54,56]. Compared to FSKD, ZSKD does not need support image with annotations, instead, using the text prompt to instruct the model to detect the specified keypoints in query image. In this case, the support set is text prompt. Though impressive for the pioneering FSKD and ZSKD models, we identify an important issue limiting the progress: *the prompt diversity for current keypoint detection methods is far from opened, especially in the aspects of modality, semantics, and language.*

Modality diversity. Most existing keypoint detection models cannot support multimodal prompts, *e.g.*, image, text, or both (see Fig. 1). Multimodal prompting is more friendly in real-world interaction, and coherent with the human concept recognition. We not only see the objects, but also describe objects with language. Our work extends existing FSKD and ZSKD, building a more general zero- and few-shot keypoint detection by leveraging a multimodal prototype set and aligning the visual keypoint features towards the textual features. While straightforward, our method enables one to study the advantages of respective modality data and exploit them for better model training and testing.

Semantic diversity. Considering keypoints between seen and unseen species, there exist large similarity yet difference in semantics. A strong advantage of text prompt is that the keypoints with the same semantics share high similarity in language across species, which enables excellent ZSKD on base keypoints. However, if the text has different semantics, significant domain gap arises. For instance, "the eye of a cat" *vs*. "the eye of a dog" has cosine similarity of 0.93 using CLIP

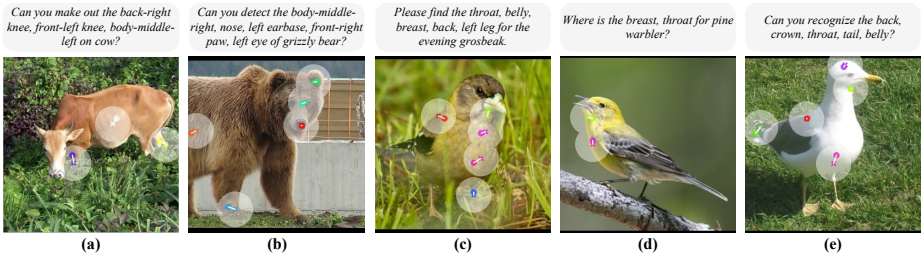


Fig. 2: Examples of keypoint detection under diverse text prompting. With LLM, our method can deal with diverse texts, showing potential for real-world applications. The circles and crosses refer to predictions and GT, respectively. A keypoint is regarded as a correct detection if falling in the white shadow area that signifies PCK@0.1.

text embeddings, whereas for "the eye of a cat" *vs*. "the knee of a dog" it drops to 0.77. We observe that the keypoint detectors cannot perform well if text is unseen during training. To bridge the gap, we propose to open the semantic diversity by adding auxiliary keypoints and texts into training, where the auxiliary texts are reasoned by LLM given the base keypoints. To ease the reasoning error, we ask LLM a good question using chain of thought [48]. We also propose a novel sampling strategy with false text control to improve the matching between auxiliary keypoints and texts. With them together, our model significantly improves ZSKD on novel keypoints.

Language diversity. Following zero-shot image classification by CLIP [35], existing ZSKD also constructs *simple text prompts* based on templates, *e.g.*, "⟨keypoint⟩" or "the ⟨keypoint⟩ of a ⟨category⟩ in the photo", where ⟨·⟩ is replaced by class names. In real-world interaction, humans tend to question in diverse styles, which results in *diverse text prompts*. For example, "Please locate the right-back leg on cat", "Can you find the left eye, right ear and nose in image?", *etc*. A natural question is how to handle the diverse text prompting? To address this, we propose a simple yet effective approach, whose key idea is to borrow the large language model (LLM) such as GPT3.5 [4] or Vicuna [9, 58] to parse the diverse texts via prompt engineering. After extracting the keypoint type and object category, they will be synthesized into simple text prompts to instruct the detection model. As such, the diverse text prompting can be transformed into simple text prompting, which sheds light on opening language diversity (Fig. 2).

To summarize, in this paper, we propose an OpenKD model with several intriguing features: 1) supporting both visual and textual prompts, 2) having the potential to handle unseen texts and diverse texts, and 3) maintaining strong generality and performance on ZSKD and FSKD. We report that LLM is capable of being a reasoner for text interpolation, and a good language parser for parsing diverse texts. We also contribute four diverse text prompt sets for the popular Animal pose dataset [5], AwA [2], CUB [45], and NABird [42] for fair evaluations. To our best knowledge, we are the first to open semantics and language diversity of text prompts for keypoint detection.

2 Related Work

Keypoint Detection has been widely studied ranging from the traditional interest points [10, 25] to deep corner detection [57], semi-supervised [18, 31, 46] and fully-supervised keypoint estimation [6, 8, 12, 33, 37, 40, 51, 52]. There are two major classes of methods for deep keypoint localization: i) direct coordinates based regression [7, 41] and ii) heatmap based regression. In contrast to existing heatmap based models dedicated to specific body parts, *e.g.*, top-down [12,16,37] and bottom-up pose estimators [6, 8, 33], our OpenKD model offers more flexible keypoint detection, breaking the limitation of keypoint types to be detected.
Few-shot Keypoint Detection is more versatile and data-efficient compared to the supervised paradigms. Similar to standard few-shot learning (FSL) [44], FSKD also takes episodes for training and evaluation. Many well-known FSL methods such as ProtoNet [36], RelationNet [38], LwoF [15] and MAML [13] have been extended to the field of keypoint detection [3, 14, 27] and serve as baselines. Recently, Lu and Koniusz [27] formalize the comprehensive FSKD settings and model the localization uncertainty of keypoints. Further, they propose a lightweight FSKD model [28] and explore FSKD under transductive setting and occlusions [29]. Bohdal *et al.* [3] propose a dataset-of-datasets, benchmarking the FSL algorithms universal to various vision tasks. FSKD also inspires class-agnostic animal pose estimation [27, 28, 50]. Compared to existing FSKD, our work pushes further by opening the text prompting, which enables more versatile keypoint detection. Moreover, text prompts can complement visual prompts, delivering better performance of detecting base keypoints on unseen species.
Zero-shot Keypoint Detection becomes feasible thanks to the great progress of vision-language pre-training (*e.g.*, CLIP [35] and BLIP [22]). Since CLIP was pre-trained on internet-scale image-text pairs that contain more semantic concepts than other datasets, CLIP can provide semantically rich features for various downstream tasks [32, 56, 59], showing strong zero-shot transfer ability.

Recently, CLAMP [56] adapts CLIP for animal pose estimation in feature-aware and spatial-aware aspects; Approach [54] also employs CLIP as backbone and proposes language-driven keypoint detection. Compared to FSKD which still requires one or more annotations, ZSKD merely demands cheap language descriptions, which make it intriguing for keypoint detection. Nevertheless, most existing ZSKD works cannot support multimodal prompts and their prompt diversity is limited. To address these issues, we propose an OpenKD model which further opens modality, semantics, and language. To handle unseen texts, we propose to generate auxiliary keypoint-text pairs which significantly help reason novel keypoints on unseen species. Moreover, we contribute a simple yet effective language parsing module to handle diverse text prompts.
Foundation Models include a broad set of models pre-trained over large-scale datasets, including LLMs (*e.g.*, GPT3.5 [4]), VLMs (*e.g.*, CLIP [35]), and Multimodal models (*e.g.*, GPT4 [1]). Transferring the knowledge from these foundation models to specific tasks *in a cost-effective way* is popular in computer vision [19,24,26,54,56]. Thus, we follow this trend and leverage LLM like GPT3.5 (relatively cheap) to perform reasoning and parsing for keypoints.

3 Method

3.1 Model Architecture

Following the general few-shot learning [36, 38, 44], Z-FSKD model is evaluated on episodes, each of which includes a support set and a query set. The query set is comprised of query images, while the support set gives the prompts. If the *visual prompts*, *i.e.*, K support images with keypoint annotations ($K \geq 1$), are given, then the problem is defined as *K-shot detection*. If only the *text prompt* (*i.e.*, the language description) is given, then $K = 0$ thus the problem becomes *zero-shot detection*. The goal of Z-FSKD is to detect the corresponding keypoints in query image given the prompts, whether visual prompt, text prompt, or both. Such an approach allows the model to effectively respond to diverse inputs.

The overview of our model inference is shown in Fig. 3, which mainly includes four stages: i) image/text feature extraction, ii) feature adaptation, iii) keypoint prototype set building, iv) correlation, decoding and heatamp fusion. The overview of our model training is shown in Fig. 4, which includes the novel approaches for performance improvement.

Image/Text Feature Extraction Since the input of our model involves vision and language two different modalities, we wish the features extracted from image and text have smaller modality gap, so that the text can somewhat correlate with image regions of the same semantics. Such a property helps the model find keypoint locations in query image. To this end, we resort to CLIP [35] pre-trained on large-scale image-text pairs. The CLIP includes image and text encoders. The text encoder is generally a transformer [43] while the image encoder can be CNN [21] or ViT [11]. We empirically found that the CLIP image encoder based on RN50 [17, 35], a CNN model, gives high efficiency in consideration of both performance and cost. Thus, we choose it as the default backbone. When extracting the image features, original CLIP image encoder only retains the classification token $\mathbf{x}_{\text{class}}$ and discards image tokens $\mathbf{X}$ by *attention pooling* as

$$\mathbf{x}_{\text{class}} = \text{AvgPool}(\mathbf{X}), \quad \mathbf{x}'_{\text{class}} = \text{Attention}(\mathbf{x}_{\text{class}}\mathbf{W}_q, \mathbf{X}\mathbf{W}_k, \mathbf{X}\mathbf{W}_v)\mathbf{W}_o. \tag{1}$$

While it is natural for image-text matching, we require the image tokens to recover the spatial locations of keypoints. Thus, we propose to obtain the image tokens via a projection composed by $\mathbf{W}_v$ and $\mathbf{W}_o$ as

$$\mathbf{X}' = \mathbf{X}\mathbf{W}_v\mathbf{W}_o. \tag{2}$$

In this way, we can not only handle the *channel inconsistency* between raw image tokens $\mathbf{X}$ and text features, but also *reuse* the projection layers $\mathbf{W}_v$ and $\mathbf{W}_o$.

For brevity, we assume that per input episode contains one query image $\mathbf{I}^{\text{q}}$, one support image $\mathbf{I}^{\text{s}}$ with N annotated keypoints, and N keypoint texts. With Eq. 2, both the support and query images can be encoded and projected as the support and query feature maps as $\mathbf{X}^{\text{s}} = \mathcal{F}_{\text{v}}(\mathbf{I}^{\text{s}})$ and $\mathbf{X}^{\text{q}} = \mathcal{F}_{\text{v}}(\mathbf{I}^{\text{q}})$ in deep feature space $\mathbb{R}^{l\times l\times d}$ via a shared CLIP image encoder $\mathcal{F}_{\text{v}}$. Further, with the CLIP text encoder $\mathcal{F}_{\text{t}}$, the N texts are firstly tokenized, then encoded as the text features $\mathbf{t}_n \in \mathbb{R}^{m\times d}$, where $n = 1, 2, \cdots, N$; and m is sequence length.

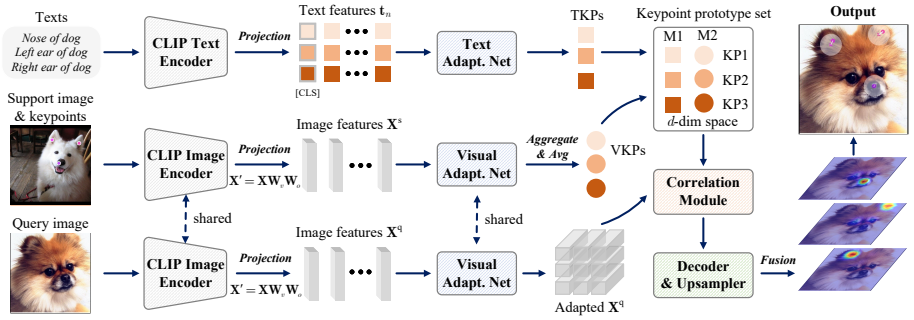


Fig. 3: The sketch of model inference. Our OpenKD allows testing under visual prompt, text prompt, or both. For clarification, we show the "both" case (*i.e.*, 1-shot with text testing). We firstly extract the deep features of texts, support and query images via CLIP, and then adapt both modalities of features via residual refinement. After extracting the visual keypoint prototype (VKP) and textual counterpart, we build the prototype set to perform class-agnostic correlation and heatmap decoding. Finally, we fuse the heatmaps induced by two modalities (*i.e.*, M1 & M2) to obtain predictions.

Feature Adaptation As CLIP provides general yet transferrable features, we adapt it to a new multimodal feature space suitable for keypoint detection. Inspired by residual learning [17], we propose two adaptation nets $\mathcal{A}_\text{v}$ and $\mathcal{A}_\text{t}$ to respectively refine the obtained image and text features in a residual way:

$$\mathbf{X}^\text{s} := \mathbf{X}^\text{s} + \mathcal{A}_\text{v}(\mathbf{X}^\text{s}), \quad \mathbf{X}^\text{q} := \mathbf{X}^\text{q} + \mathcal{A}_\text{v}(\mathbf{X}^\text{q}), \quad \mathbf{t}_n := \mathbf{t}_n + \mathcal{A}_\text{t}(\mathbf{t}_n). \tag{3}$$

The adaptation net $\mathcal{A}$ is highly scalable and can be based on transformer [43] or bottleneck [17] as refinement blocks. We found that $\mathcal{A}$ can adapt features well.

Keypoint Prototype Set Once we obtain the adapted features of support image and text prompts, *i.e.*, $\mathbf{X}^\text{s}$ and $\mathbf{t}_n$, we propose to convert them into keypoint prototypes, which are unified in a keypoint prototype set.

For visual prompt, recall that keypoint labels are provided along with the support image. Thus, one can aggregate the the local features for each support keypoint $\mathbf{p}_n$ by using pixelwise weighted summation between feature map $\mathbf{X}^\text{s}$ and Gaussian heatmap $\mathbf{H}(\mathbf{p}_n; \sigma)$, yielding the visual keypoint representation (VKR) as $\mathbf{\Phi}_n \in \mathbb{R}^d$. The σ is the standard deviation that controls the Gaussian spread. If K support images are given, namely in few-shot case, the VKRs of the same type of keypoints $\mathbf{\Phi}_{k,n}$ will be averaged to build the visual keypoint prototype (**VKP**) $\mathbf{\Psi}^\text{v}_n = \frac{1}{K}\sum_k \mathbf{\Phi}_{k,n}$. Analogously, one can develop textual keypoint prototype (**TKP**) $\mathbf{\Psi}^\text{t}_n$ by text features if provided with multiple texts per keypoint. Finally, one can build the keypoint prototype set as $\mathcal{T} = \mathcal{T}^\text{v} \cup \mathcal{T}^\text{t}$, where $\mathcal{T}^\text{v} = \{\mathbf{\Psi}^\text{v}_n\}$ is the VKPs and $\mathcal{T}^\text{t} = \{\mathbf{\Psi}^\text{t}_n\}$ is the TKPs. In this way, prompts of different modalities can be summarized as keypoint prototypes in a shared d dimensional feature space, which enables our model to flexibly handle various modalities. Moreover, each prototype in set $\mathcal{T}$ can guide the model to induce a keypoint heatmap, establishing the zero- and few-shot keypoint detection.

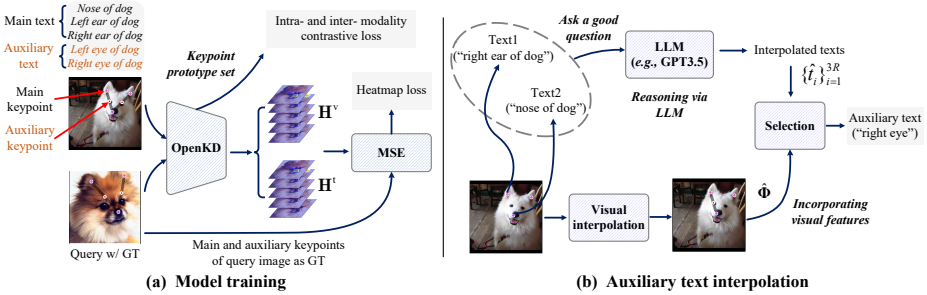


Fig. 4: Model training and text interpolation. (a) In addition to multi-group heatmap regression, we improve model performance by introducing intra- and inter-modality contrastive learning and the novel auxiliary keypoint and text learning. (b) We exploit LLM for auxiliary texts interpolation and explore incorporating visual keypoint features for text selection in order to mitigate the noisy texts or false texts.

Correlation, Decoding and Fusion To discover the corresponding keypoints in query image $\mathbf{I}^{\mathrm{q}}$, each keypoint prototype $\mathbf{\Psi}_n$, either VKP $\mathbf{\Psi}_n^{\mathrm{v}}$ or TKP $\mathbf{\Psi}_n^{\mathrm{t}}$, is required to be correlated with the query feature map $\mathbf{X}^{\mathrm{q}}$. To this end, we leverage a correlation module $\mathcal{C}$ which takes both $\mathbf{\Psi}_n$ and $\mathbf{X}^{\mathrm{q}}$ as input to produce the attentive feature maps $\mathbf{A}_n = \mathcal{C}(\mathbf{\Psi}_n, \mathbf{X}^{\mathrm{q}})$, where $\mathbf{A}_n \in \mathbb{R}^{l\times l\times d}$ and $n = 1, 2, \cdots, N$. For the correlation module $\mathcal{C}$, we explored multiple variants and found that the simple *cross correlation* already yields high performance, *i.e.*, $\mathbf{A}_n = \mathbf{X}^{\mathrm{q}} \odot \mathbf{\Psi}_n$, where $\odot$ denotes the channel-wise multiplication.

Subsequently, a *class-agnostic* decoder $\mathcal{D}$ is devised to convert each attentive feature map into a keypoint heatmap, *i.e.*, $\mathbf{H}_n = \mathcal{D}(\mathbf{A}_n)$, where $\mathbf{H}_n \in \mathbb{R}^{l\times l}$. For heatmap regression based keypoint localization, the higher resolution of heatmap can greatly reduce the coordinate decoding error. Thus, a lightweight upsampling module $\mathcal{U}$ is adopted to further refine heatmaps, *i.e.*, $\mathbf{H}_n := \mathcal{U}(\mathbf{H}_n)$.

Since each modality of prototypes can induce one group of heatmaps, considering visual and textual modalities, we have two groups of heatmaps $\mathbf{H}^{\mathrm{v}} = \{\mathbf{H}_n^{\mathrm{v}}\}_{n=1}^N$ and $\mathbf{H}^{\mathrm{t}} = \{\mathbf{H}_n^{\mathrm{t}}\}_{n=1}^N$. In testing phase, we *fuse* the upsampled multi-group heatmaps as the final output $\mathbf{H} \in \mathbb{R}^{N\times ul\times ul}$ (u is upsampling factor):

$$\mathbf{H} = (\mathbf{H}^{\mathrm{v}} + \mathbf{H}^{\mathrm{t}})/2. \tag{4}$$

During training phase, inspired by HigherHRNet [8], we perform *group-specific supervision* for heatmaps as

$$\mathcal{L}_{\mathrm{kp}} = (\|\mathbf{H}^{\mathrm{v}} - \mathbf{H}^{*}\|_2^2 + \|\mathbf{H}^{\mathrm{t}} - \mathbf{H}^{*}\|_2^2)/2, \tag{5}$$

where $\mathbf{H}^{*}$ denotes the groundtruth heatmap that encodes the query keypoint.

Improving Model by Contrastive Learning Since the keypoint prototype directly impacts performance, we aim for its representativeness to be as good as possible. The same type of keypoints should have higher invariance across species, *e.g.*, the ear of cat, dog, and cow, while different types are required to

be sufficiently discriminative such that our model can distinguish the ambiguous texts or fine-grained texts such as "the left ear" *vs.* "the right ear". To this end, we propose to introduce a contrative loss over TKPs. If randomly sampling *two* species (s, s') at a time and each species pertaining to an episode, we have pairwise sets of TKPs, *i.e.*, $\mathcal{T}_s^{\mathrm{t}} = \{\boldsymbol{\Psi}_n\}$ and $\mathcal{T}_{s'}^{\mathrm{t}} = \{\boldsymbol{\Psi}'_n\}$, which can form similarity matrix as

$$\mathbf{J}(\mathcal{T}_s^{\mathrm{t}}, \mathcal{T}_{s'}^{\mathrm{t}}) = \begin{pmatrix} \cos(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}'_1) & \cdots & \cos(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}'_N) \\ \vdots & & \vdots \\ \cos(\boldsymbol{\Psi}_N, \boldsymbol{\Psi}'_1) & \cdots & \cos(\boldsymbol{\Psi}_N, \boldsymbol{\Psi}'_N) \end{pmatrix}, \tag{6}$$

where $\cos(\cdot, \cdot)$ denotes the cosine similarity. Then, the contrastive loss within textual modality $\mathcal{L}_{\mathrm{tt}}$ becomes

$$\mathcal{L}_{\mathrm{tt}}^{s \to s'} = -\langle \mathbb{I}, \log(\mathrm{softmax}(\mathbf{J}(\mathcal{T}_s^{\mathrm{t}}, \mathcal{T}_{s'}^{\mathrm{t}})/\tau)) \rangle, \quad \mathcal{L}_{\mathrm{tt}} = \frac{1}{2}(\mathcal{L}_{\mathrm{tt}}^{s \to s'} + \mathcal{L}_{\mathrm{tt}}^{s' \to s}), \tag{7}$$

where $\mathbb{I}$ is the identity matrix and $\langle \cdot, \cdot \rangle$ denotes inner product. Since CLIP is trained via *image-level alignment* instead of *keypoint-level alignment* with text, we wish VKPs and TKPs were aligned so that we can exploit their respective advantages to help each other. This motivates us to propose the second contrastive loss between VKPs and TKPs. Similar to Eq. 7, we build $\mathcal{L}_{\mathrm{vt}}$. We find that the visual-textual alignment improves 1-shot testing, but unintentionally degrades 0-shot performance slightly. We discover the principle behind this is better clustering effects of textual keypoint representations than visual ones. To address this issue, we perform *stop gradient* on TKPs in $\mathcal{L}_{\mathrm{vt}}$, thus enforcing the VKPs to align towards TKPs. We also find adding $\mathcal{L}_{\mathrm{vv}}$ does not further improve scores. We conjecture that after aligning VKPs to TKPs via $\mathcal{L}_{\mathrm{vt}}$, the visual features will also improve discrimination along with textual features as $\mathcal{L}_{\mathrm{tt}}$ is applied.

Optimization By incorporating the heatmap regression loss $\mathcal{L}_{\mathrm{kp}}$, intra- and inter-modality contrastive loss $\mathcal{L}_{\mathrm{tt}}$ and $\mathcal{L}_{\mathrm{vt}}$, we have the overall loss as $\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{kp}} + \lambda_2 \mathcal{L}_{\mathrm{tt}} + \lambda_3 \mathcal{L}_{\mathrm{vt}}$. We set the loss weight $\lambda_1 = 1$, $\lambda_2 = \lambda_3 = 0.002$.

3.2 Training with Auxiliary Keypoints and Texts

Localizing the seen keypoints and texts is relatively easy to handle, but it becomes difficult if the unseen ones are encountered. Thus, we propose to add auxiliary keypoints and texts interpolated from visual and textual domains into training, which significantly improves the ability to detect novel keypoints.

Visual Interpolation Following [27], we generate the auxiliary keypoints in visual domain via interpolation $\mathcal{I}^{\mathrm{v}}(z; \mathbf{p}_1, \mathbf{p}_2)$, where $z \in (0, 1)$ is the interpolation node and $\mathbf{p}_i \in \mathbb{R}^2$ are the end points that form an interpolation path. An off-the-shelf saliency detector [49] is adopted to filter the auxiliary keypoints outside of the foreground [27]. Via interpolation $\mathcal{I}^{\mathrm{v}}$ in the visual domain, the auxiliary keypoints $\hat{\mathbf{p}}$ can boost the visual diversity of training keypoints.

Text Interpolation and Selection The $\mathcal{I}^{\mathrm{v}}$ can assist the model in handling unseen keypoints in visual prompt, but it helps little if prompting unseen texts.

Consequently, we propose using text interpolation $\mathcal{I}^{\mathrm{t}}(z; t(\mathbf{p}_1), t(\mathbf{p}_2), c)$ to mitigate this issue, where $t(\mathbf{p}_i)$ are the names of end points $\mathbf{p}_i$ and c is the object category. Thanks to great language processing ability of LLM (*e.g.*, GPT3.5 [4]), we propose to ask LLM a good question to reason the auxiliary keypoint texts given base keypoints (Fig. 4(b)), *e.g.*, to infer the keypoint text between "thigh" and "paw" of a "cat", we can design a vanilla prompt for the LLM as:

"Please give me three most common body parts/keypoints at 1/2 between thigh and paw of a cat. Please answer in concise words."

Apparently, we expect that the answers returned by LLM should include "knee" or "elbow". In practice, the keypoint types are usually more complex thus the reasoning is challenging. To enhance reasoning, we improve the above vanilla prompt into a more advanced one by using chain of thought (CoT) [48], whose key idea is to firstly provide an example with analysis and then ask the question. We provide the detailed prompts and examples in **§A of Suppl.**

Considering the LLM suffers from some randomness and might produce erroneous answers, we let LLM give three most possible answers for each interpolation path and repeat this step R times, yielding a text pool with $3R$ texts, *i.e.*, $\{\hat{t}_i\}_{i=1}^{3R}$. As such, the GT text has higher chance to be included. Afterwards, we need to select one interpolated text from the pool to form a pair with a visually interpolated keypoint. To this end, a simple yet effective approach is to sample from top-1 results of all repetitions. However, it does not use the visual features for selection. An interesting approach is to use the correlation between visual and text features for selection as follows:

$$j = \arg\max_i \cos(\hat{\mathbf{\Phi}}, \hat{\mathbf{t}}_i), \quad \hat{t}^* = \hat{t}_j, \tag{8}$$

where $\hat{\mathbf{\Phi}}$ is the auxiliary keypoint feature of $\hat{\mathbf{p}}$, $\hat{\mathbf{t}}_i$ is text feature of $\hat{t}_i$, and $\hat{t}^*$ is the selected text. However, we found that sometimes it leads to performance drop. We suspect it is because CLIP's keypoint-text features are not fine-grained aligned enough on those regions beyond annotations. Opposed to picking, we observe such a correlation could be used to reject low-quality text if keypoint-text similarity $\cos(\hat{\mathbf{\Phi}}, \hat{\mathbf{t}}_i)$ is low. Consequently, we propose a more advanced sampling with false text control (**FTC**). Namely, we sample the text within the range top-η but the text will be rejected if the similarity $\cos(\hat{\mathbf{\Phi}}, \hat{\mathbf{t}}_i)$ is below the threshold α. In this way, we can form auxiliary keypoint-text pairs with higher quality to enhance model training. We apply $\mathcal{I}^{\mathrm{v}}$ to both support and query images to generate support and query auxiliary keypoints $\hat{\mathbf{p}}^{\mathrm{s}}$ and $\hat{\mathbf{p}}^{\mathrm{q}}$, respectively. As in Eq. 5, $\hat{\mathbf{p}}^{\mathrm{q}}$ encodes GT heatmap supervising the heatmaps induced by $\hat{\mathbf{p}}^{\mathrm{s}}$ and $\hat{t}^*$.

3.3 LLM is a Good Language Parser

To handle *diverse text prompts* in keypoint detection, unlike current fashion of utilizing LLMs for text generation, we use LLM for text decomposition. For example, if our model receives a diverse text prompt like "Can you localize the left eye and nose of cat?", we firstly leverage an LLM (*e.g.*, GPT3.5 [4]) to perform text parsing by providing it with a simple yet effective prompt as follows:

Please extract the animal and keypoint keywords from the below text: "Can you localize the left eye and nose of cat?"
Then, LLM returns the parsed texts regarding ⟨keypoint⟩ and ⟨object⟩, *i.e.*, the "left eye", "nose", and "cat" in the example. We directly utilize parsed results to synthesize *simple text prompts* to instruct our model and predict all the keypoints corresponding to the text. The problem of diverse text prompting becomes simple text prompting, thus opening the language diversity for keypoint detection.

To fairly evaluate our approach, we manually collect 100 diverse text templates which cover most scenarios for questions (**§B of Suppl.**). Take one as an example: "Where is the ⟨keypoint⟩ for ⟨object⟩?" Afterwards, we develop four diverse text prompt sets for the popular Animal pose dataset [5], AwA [2], CUB [45], and NABird [42]. Specifically, we accompany each object instance with a diverse text prompt synthesized by randomly sampling one template and *one to N valid keypoints*. Compared to traditional text prompts which only have N variants, our constructed text prompt sets are more diverse, whose space could be as large as $100 \cdot (C_N^1 + C_N^2 + \cdots + C_N^N) = 100 \cdot 2^N$ variants. For example, as CUB has 11 keypoints, its space is larger than 10^5. We randomly sample 1000 diverse texts from each dataset and evaluate the LLM's parsing performance. Table 5b reports that GPT3.5 has over 96% accuracy in parsing keypoints from text, which shows that LLM can be a good language parser.

4 Experiments

4.1 Experimental Settings

Datasets and Splits. 1) Animal pose dataset [5] has five mammal species *cat*, *dog*, *cow*, *horse*, and *sheep*. Each animal species is alternately chosen as unseen species for testing while the remaining four as seen species for training, which yields *five* subproblems; **2)** AwA [2] has 35 diverse animal species with 10064 images. For AwA, 25 species are for training and 10 for testing; **3)** CUB [45] consists of 200 bird species with 15 keypoint annotations. We use 100 species for training, 50 for validation, and 50 for testing; **4)** NABird [42] is larger than CUB and has 555 categories. The species splits are 333, 111, and 111 for training, validation, and testing. To test model generalization, we follow [27] to split keypoints into base and novel sets, and report the performance of both base and novel keypoint detection on unseen species.
Metric. The percentage of correct keypoints (PCK) is used. A predicted keypoint is correct if its distance to GT $d \leq \rho \cdot \max(w_{\text{bbx}}, h_{\text{bbx}})$, where w_{bbx} and h_{bbx} are the edges of object bounding box. Following [27], we set ρ to 0.1.
Implementation Details. The input image size for all models is 384×384. We freeze the CLIP text encoder but finetune last two layers of image encoder. The temperature τ used in contrastive loss is 0.05. Same to [27], the visual auxiliary keypoints are generated on pre-defined interpolation path, and we set $z = 0.5$. For the text interpolation, to reduce randomness, we use GPT3.5 to reason $R = 3$ times for all datasets except CUB with $R = 10$. Our model is trained with 40k episodes. We report results using 1000 test episodes.

Table 1: Main results on the Animal pose dataset. Each row shows the results tested on the model trained in one setting. Each PCK score is the average of five subproblems.

#	Training settings				1-shot testing		0-shot testing		1-shot w/ text	
	Kp	*Aux. kp*	*Text*	*Aux. text*	Novel	Base	Novel	Base	Novel	Base
1	✓				21.36	50.84	1.26	1.93	16.16	32.81
2	✓	✓			47.54	49.45	2.00	2.10	39.24	34.11
3			✓		16.18	35.46	25.60	61.64	27.02	61.21
4			✓	✓	15.87	31.57	63.14	65.31	**63.30**	**65.52**
5	✓		✓		21.46	52.15	22.42	61.07	23.88	60.49
6	✓	✓	✓	✓	**50.32**	**54.39**	**63.37**	**65.59**	63.19	64.93

Compared Methods. For few-shot keypoint detection (FSKD), as in previous works, we compare the few-shot learning models such as *ProtoNet* [36], *RelationNet* [38], and *LwoF* [15], and FSKD-dedicated works *FSKD-R/-D* [27]. For zero-shot keypoint detection (ZSKD), we adopt the source code of CLAMP [56] and compare it in our experiments for fairness. Moreover, we build a *Baseline* sharing same backbone with our model but we do not use the specifically proposed auxiliary texts, contrastive loss, *etc*. We denote our method as *OpenKD*.

4.2 Main Results

Firstly, we explore *single-modal training* and the impacts of whether or not to add the auxiliary keypoints/texts (1st-4th row, Table 1). When only using main keypoints (1st row) or main texts (3rd row), one can observe that the models perform well on base keypoints in 1-shot (50.84%) or 0-shot (61.64%), but fail to detect novel keypoints. However, if the model is trained by adding the auxiliary keypoints (2nd row) or auxiliary texts (4th row), one can observe the remarkable improvements on novel keypoint detection, *e.g.*, 47.54% *vs*. 21.36% in 1-shot testing for the model trained by visual prompts (2nd row *vs*. 1st row); and also 63.14% *vs*. 25.60% in 0-shot testing for the model trained by text prompts (4th row *vs*. 3rd row). The significant gains confirm the benefits of visual interpolation [27] and also highlight the effectiveness of our proposed textual interpolation. **Secondly**, we investigate *multimodal training* (5th-6th row, Table 1). As one can see, our model performs well under multimodal training (6th row), and greatly outperforms the model trained by multimodal prompts without auxiliary keypoints and texts (5th row). The higher performance is due to the fact that auxiliary keypoints and texts could boost the visual and textual semantic diversity, thus enabling novel keypoint detection under 1-shot and 0-shot setting. **Thirdly**, we observe that 0-shot testing on base keypoints is much higher than 1-shot, which shows that the texts can be an excellent representation to guide keypoint detection. **Lastly**, we examine *multimodal testing*, *i.e.*, performing *1-shot with text* testing, our model obtains 63.19% and 64.93% on novel and base keypoints, which highlights that our model strongly combines the advantages of both modalities, mitigating the weakness of individual modality.

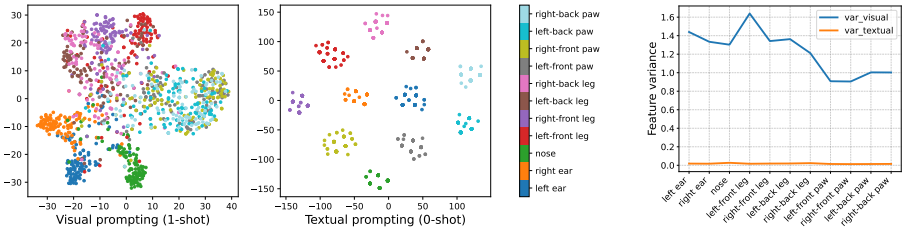
Table 2: Results on 1-shot keypoint detection for unseen species. The PCK scores on novel and base keypoint detection are reported. +**T** means adding texts during testing.

Setting	Model	Animal Pose Dataset						AwA	CUB	NAB
		Cat	Dog	Cow	Horse	Sheep	Avg			
Novel	ProtoNet	19.68	16.18	14.39	12.05	15.06	15.47	29.57	51.32	36.65
	RelationNet	22.15	17.19	15.47	13.58	16.55	16.99	20.91	56.59	34.02
	LwoF	22.47	19.39	16.82	16.40	16.94	18.40	28.54	54.75	34.19
	FSKD-R	46.05	40.66	37.55	38.09	31.50	38.77	51.81	77.90	54.01
	FSKD-D	52.36	47.94	44.07	42.77	36.60	44.75	64.76	77.89	**56.04**
	OpenKD	60.36	53.58	47.59	49.01	41.05	50.32	66.71	**78.39**	53.35
	OpenKD+**T**	**69.26**	**66.81**	**62.40**	**63.21**	**54.27**	**63.19**	**79.02**	73.29	53.40
Base	ProtoNet	45.80	39.83	34.88	35.80	32.33	37.73	57.17	80.36	73.18
	RelationNet	51.03	45.85	39.86	41.97	37.19	43.18	57.31	79.40	78.85
	LwoF	50.05	44.64	43.47	43.35	37.84	43.87	63.87	81.96	81.39
	FSKD-R	57.12	51.12	47.83	49.71	43.71	49.90	65.26	87.94	87.84
	FSKD-D	56.38	51.29	48.24	49.77	43.95	49.93	66.39	87.71	86.99
	OpenKD	63.61	55.43	51.18	53.87	47.86	54.39	74.22	87.45	85.11
	OpenKD+**T**	**70.47**	**65.09**	**62.84**	**66.46**	**59.81**	**64.93**	**84.50**	**91.81**	**91.23**

Table 3: Results on the 0-shot keypoint detection for unseen species. CLAMP† is the variant trained by adding our interpolated auxiliary texts.

Setting	Model	Animal Pose Dataset						AwA	CUB	NAB
		Cat	Dog	Cow	Horse	Sheep	Avg			
Novel	Baseline	25.64	24.29	25.29	18.14	25.51	23.77	28.95	34.94	30.00
	CLAMP	20.90	24.06	27.07	16.86	20.73	21.92	38.96	37.09	18.18
	CLAMP†	61.70	58.09	**61.42**	64.86	53.13	59.84	77.66	69.30	50.81
	OpenKD	**71.07**	**66.49**	60.30	**65.53**	**53.45**	**63.37**	**78.64**	**70.16**	**52.29**
Base	Baseline	58.69	56.62	56.58	60.00	51.35	56.65	73.45	87.12	71.25
	CLAMP	60.71	54.04	62.58	62.73	57.30	59.47	84.16	90.97	79.10
	CLAMP†	60.30	56.49	62.69	60.90	57.16	59.51	83.76	90.65	83.65
	OpenKD	**71.34**	**66.60**	**64.53**	**67.26**	**58.23**	**65.59**	**84.32**	**91.72**	**90.63**

Comparisons on FSKD and ZSKD: We comprehensively conduct the few-shot keypoint detection across four datasets. Table 2 shows our OpenKD model significantly outperforms compared methods in 15 out of 18 tasks in detecting novel or base keypoints. Moreover, if adding the texts during testing, the performance of our model further improves, yielding 63.19% on the Animal pose dataset (63.19% *vs*. 44.75% of prior-art FSKD-D) and 79.02% on AwA. We also notice the improvements on CUB and NAB are modest, which might be due to the auxiliary texts being relatively harder to reason, thus bringing lesser benefit to few-shot model. For ZSKD (Table 3), again, our model outperforms other



(a) (b)

Fig. 5: Different clustering effects of two modalities on base keypoints in unseen species (a) and statistical feature variance per keypoint (b).

Table 4: Ablation study. (a) Intra- and inter-modality contrastive loss. $\mathcal{L}^*_{\text{vt}}$ denotes no stop gradient on text features; (b) Text interpolation. Baseline†: w/o using auxiliary texts; CoT: chain of thought; Corr: Eq. 8; FTC: sampling with false text control (ours).

AwA	1-shot		0-shot	
	Novel	Base	Novel	Base
1: w/o CL	65.56	72.17	76.71	81.67
2: $\mathcal{L}_{\text{tt}}$	66.05	72.35	78.28	84.00
3: $\mathcal{L}_{\text{tt}}$+$\mathcal{L}^*_{\text{vt}}$	66.70	73.91	77.10	83.85
4: $\mathcal{L}_{\text{tt}}$+$\mathcal{L}_{\text{vt}}$	**66.71**	**74.22**	**78.64**	**84.32**
5: $\mathcal{L}_{\text{tt}}$+$\mathcal{L}_{\text{vt}}$+$\mathcal{L}_{\text{vv}}$	66.27	74.08	78.60	83.42

(a)

0-shot testing	AwA		CUB	
	Novel	Base	Novel	Base
1: Baseline† (▲)	34.31	83.53	36.89	88.85
2: ▲+w/o CoT	73.80	82.79	68.03	88.93
3: ▲+CoT	78.30	**84.32**	68.68	88.98
4: ▲+CoT+Corr	62.12	83.13	**71.80**	**94.27**
5: ▲+CoT+FTC	**78.64**	**84.32**	70.16	91.72

(b)

methods (in 17 out of 18 tasks). The improvement of CLAMP† on novel keypoint detection strongly shows the benefits of adding auxiliary texts into training.

4.3 Model and Performance Analysis

Why 0-shot can outperform 1-shot? One may observe the 0-shot testing scores on base keypoints are always higher than 1-shot testing. In this case, the training texts are accurate (*cf*. auxiliary text is noisy) and have high semantic similarity to testing texts. We dig more to visualize the corresponding text and keypoint features. Fig. 5 shows that textual keypoint features enjoy better *clustering effects* and lower variance than visual ones, which suggests textual representations of base keypoints are better thus 0-shot yields higher scores.

Contrastive Learning (CL). After applying $\mathcal{L}_{\text{tt}}$, we observe the 0-shot performance significantly improves, *e.g.* 84.00% *vs*. 81.67% (2nd row, Table 4a), as CL improves the discrimination between text features. However, when adding visual-textual alignment (3rd row), 1-shot testing scores boost but slightly decreasing 0-shot scores. We suspect it is due to the imbalance of representation quality between multimodal features. After adding stop gradient, it mitigates negative impacts from "weak" modality, thus achieving higher scores (4th row). We also observe $\mathcal{L}_{\text{vv}}$ does not further improve scores in our losses (5th row).

Text Interpolation. Table 4b shows that using chain of thought (CoT) significantly improves performance, especially on AwA (3rd row), as CoT improves the

Table 5: (a) Study on repetition R; (b) Parsing results with GPT3.5 and Vicuna.

R	AwA		CUB	
	Novel	Base	Novel	Base
1	67.34	83.98	64.21	91.40
3	78.64 (+11.3)	84.32	69.81 (+5.60)	92.38
5	76.36 (+9.02)	83.73	69.81 (+5.60)	92.02
10	76.61 (+9.27)	83.36	70.16 (+5.95)	91.72

(a)

Dataset	GPT3.5		Vicuna	
	Acc. kp	*Acc. obj*	*Acc. kp*	*Acc. obj*
Animal	0.97	0.99	0.94	0.98
AwA	0.96	1.00	0.93	0.98
CUB	0.97	0.96	0.94	1.00
NABird	0.97	0.99	0.95	1.00

(b)

Table 6: Diverse text prompting evaluation. Results are the average of three runs.

Method	Animal Pose		AwA		CUB		NAB		Avg Drop	
	Novel	Base	Novel	Base	Novel	Base	Novel	Base	Novel	Base
OpenKD	63.37	65.59	78.64	84.32	70.16	91.72	52.29	90.63	-	-
No parsing	13.89	15.18	13.44	21.70	22.09	17.48	14.55	17.62	50.12↓	65.07↓
Vicuna	57.54	59.84	73.97	78.18	67.89	87.39	50.45	86.59	3.65↓	5.06↓
GPT3.5	61.45	62.82	76.23	80.07	69.39	90.65	52.21	89.14	**1.29↓**	**2.39↓**

quality of reasoned texts. After applying FTC, the scores strike a high balance in all datasets (5-th row), as our FTC rejects low-quality texts and further reduces noise after including visual features into decision. We also explore the repetition R for text reasoning. Table 5a shows the importance of setting $R > 1$.

Diverse Text Prompting. To evaluate our model under diverse text prompting, we randomly sample 1000 diverse texts from each dataset to conduct zero-shot testing. As shown in Table 6, after coupling with GPT3.5, our model retains the high performance even under diverse texts (with less than 3% drop). The strong results compared to no parsing show that leveraging LLM as a parser to handle language diversity is a possible solution. Moreover, we also found coupling with GPT3.5 has higher performance compared to Vicuna, as GPT3.5 enjoys a stronger keypoint text parsing ability, *e.g.*, 96% *vs.* 93% in AwA (Table 5b).

5 Conclusion

We propose to open the prompt diversity from the aspects of modality, semantics and language, to enable a more general zero- and few-shot keypoint detection. To this end, we build a versatile OpenKD model which supports both visual and textual prompting. Moreover, to bridge the semantics gap between seen and unseen texts, we propose a novel text interpolation and the selection strategy with false text control, which significantly improves zero-shot novel keypoint detection. We also discover that LLM is a good language parser. After coupling with LLM, our model can handle diverse texts well. We hope the proposed model, text interpolation and the parsing approach will provide useful insights for versatile keypoint detection, thus we highly recommend it to vision community.

Acknowledgment

Changsheng Lu is supported by Australian Government Research Training Program (AGRTP) International Scholarship. Piotr Koniusz is supported by CSIRO's Science Digital.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Banik, P., Li, L., Dong, X.: A novel dataset for keypoint detection of quadruped animals from images. arXiv preprint arXiv:2108.13958 (2021)
3. Bohdal, O., Tian, Y., Zong, Y., Chavhan, R., Li, D., Gouk, H., Guo, L., Hospedales, T.: Meta omnium: A benchmark for general-purpose learning-to-learn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7693–7703 (2023)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
5. Cao, J., Tang, H., Fang, H.S., Shen, X., Lu, C., Tai, Y.W.: Cross-domain adaptation for animal pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9498–9507 (2019)
6. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence **43**(1), 172–186 (2019)
7. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4733–4742 (2016)
8. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5386–5395 (2020)
9. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna.lmsys.org (accessed 14 April 2023) (2023)
10. Derpanis, K.G.: The harris corner detector. York University pp. 2–3 (2004)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
12. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE international conference on computer vision. pp. 2334–2343 (2017)
13. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning. pp. 1126–1135. PMLR (2017)

14. Ge, Y., Zhang, R., Luo, P.: Metacloth: Learning unseen tasks of dense fashion landmark detection from a few samples. IEEE Transactions on Image Processing **31**, 1120–1133 (2021)
15. Gidaris, S., Komodakis, N.: Dynamic few-shot visual learning without forgetting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4367–4375 (2018)
16. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
18. Honari, S., Molchanov, P., Tyree, S., Vincent, P., Pal, C., Kautz, J.: Improving landmark localization with semi-supervised learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1546–1555 (2018)
19. Jiao, B., Liu, L., Gao, L., Wu, R., Lin, G., Wang, P., Zhang, Y.: Toward re-identifying any animal. Advances in Neural Information Processing Systems **36** (2024)
20. Koch, G., Zemel, R., Salakhutdinov, R., et al.: Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop. vol. 2. Lille (2015)
21. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
22. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
23. Li, S., Gunel, S., Ostrek, M., Ramdya, P., Fua, P., Rhodin, H.: Deformation-aware unpaired image translation for pose estimation on laboratory animals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13158–13168 (2020)
24. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
25. Lowe, G.: Sift-the scale invariant feature transform. Int. J **2**, 91–110 (2004)
26. Lu, C., Gu, C., Wu, K., Xia, S., Wang, H., Guan, X.: Deep transfer neural network using hybrid representations of domain discrepancy. Neurocomputing **409**, 60–73 (2020)
27. Lu, C., Koniusz, P.: Few-shot keypoint detection with uncertainty learning for unseen species. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19416–19426 (2022)
28. Lu, C., Koniusz, P.: Detect any keypoints: An efficient light-weight few-shot keypoint detector. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 3882–3890 (2024)
29. Lu, C., Zhu, H., Koniusz, P.: From saliency to dino: Saliency-guided vision transformer for few-shot keypoint detection. arXiv preprint arXiv:2304.03140 (2023)
30. Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M.: Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. Nature neuroscience **21**(9), 1281–1289 (2018)
31. Moskvyak, O., Maire, F., Dayoub, F., Baktashmotlagh, M.: Semi-supervised keypoint localization. arXiv preprint arXiv:2101.07988 (2021)
32. Mukhoti, J., Lin, T.Y., Poursaeed, O., Wang, R., Shah, A., Torr, P.H., Lim, S.N.: Open vocabulary semantic segmentation with patch aligned contrastive learning.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19413–19423 (2023)
33. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. pp. 483–499. Springer (2016)
34. Pereira, T.D., Aldarondo, D.E., Willmore, L., Kislin, M., Wang, S.S.H., Murthy, M., Shaevitz, J.W.: Fast animal pose estimation using deep neural networks. Nature methods **16**(1), 117–125 (2019)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
36. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. arXiv preprint arXiv:1703.05175 (2017)
37. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5693–5703 (2019)
38. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1199–1208 (2018)
39. Tang, L., Wertheimer, D., Hariharan, B.: Revisiting pose-normalization for fine-grained few-shot recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14352–14361 (2020)
40. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. Advances in neural information processing systems **27** (2014)
41. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1653–1660 (2014)
42. Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., Belongie, S.: Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 595–604 (2015)
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
44. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. Advances in neural information processing systems **29**, 3630–3638 (2016)
45. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
46. Wang, C., Jin, S., Guan, Y., Liu, W., Qian, C., Luo, P., Ouyang, W.: Pseudo-labeled auto-curriculum learning for semi-supervised keypoint localization. arXiv preprint arXiv:2201.08613 (2022)
47. Wang, L., Huynh, D.Q., Koniusz, P.: A comparative review of recent kinect-based action recognition algorithms. IEEE Transactions on Image Processing **29**, 15–28 (2019)

48. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)
49. Wu, Z., Su, L., Huang, Q.: Stacked cross refinement network for edge-aware salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7264–7273 (2019)
50. Xu, L., Jin, S., Zeng, W., Liu, W., Qian, C., Ouyang, W., Luo, P., Wang, X.: Pose for everything: Towards category-agnostic pose estimation. In: European Conference on Computer Vision. pp. 398–416. Springer (2022)
51. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose+: Vision transformer foundation model for generic body pose estimation. arXiv preprint arXiv:2212.04246 (2022)
52. Yang, S., Quan, Z., Nie, M., Yang, W.: Transpose: Keypoint localization via transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11802–11812 (2021)
53. Yao, Q., Quan, Q., Xiao, L., Kevin Zhou, S.: One-shot medical landmark detection. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. pp. 177–188. Springer (2021)
54. Zhang, H., Zhang, K., Xu, L., Lai, S., Shao, W., Zheng, N., Luo, P., Qiao, Y.: Language-driven open-vocabulary keypoint detection for animal body and face. arXiv preprint arXiv:2310.05056 (2023)
55. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based r-cnns for fine-grained category detection. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. pp. 834–849. Springer (2014)
56. Zhang, X., Wang, W., Chen, Z., Xu, Y., Zhang, J., Tao, D.: Clamp: Prompt-based contrastive learning for connecting language and animal pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23272–23281 (2023)
57. Zhao, S., Gong, M., Zhao, H., Zhang, J., Tao, D.: Deep corner. International Journal of Computer Vision pp. 1–25 (2023)
58. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena (2023)
59. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16793–16803 (2022)