GKDT: General Keypoint Detection Transformer

Changsheng Lu[1], Yuxin Chen[1], Haokun Gui[1], Rong Wang[2], Jie Yang[3], Harry Yang[1], Anton van den Hengel[4], and Jiaya Jia[1]

[1] Hong Kong University of Science and Technology, Hong Kong, China
changshengluu@gmail.com, jia@cse.ust.hk
[2] Australian National University, Canberra, Australia
[3] Tencent Inc., China
[4] Adelaide University, Adelaide, Australia
anton.vandenhengel@adelaide.edu.au

Abstract. With the emergence of various pre-trained vision and language models, computer vision is shifting from narrow-domain to open-domain recognition. The construction of a more powerful yet general keypoint detection (GKD) model to support diverse tasks has become increasingly important in the field. To this end, we firstly present a large-scale unified keypoint dataset called MegaKPT. The dataset is composed of over 1.3 million diverse object instances from twenty-nine existing datasets, and enjoys high-quality unified annotations with keypoint text descriptions. Based on MegaKPT, we develop GKDT, a simple, flexible and powerful DINOv3 based Transformer model for General Keypoint Detection. Our GKDT supports visual prompts, text prompts, or both. To enhance model training, we also propose a suite of useful strategies such as mix-modal prompted training and dynamic importance sampling. By testing over 22 test sets with seen or unseen objects, our single GKDT model shows strong performance and generality in detecting keypoints on broad categories, with most categories over 90% PCK@0.1 accuracy, offering high practical applicability to real-world problems. The dataset, models, and codes will be released at https://github.com/AlanLuSun/General-Keypoint-Detection.

1 Introduction

Keypoint detection is a fundamental task in computer vision that aims to localize a set of keypoints on an image. Compared to object detection, keypoint detection provides more fine-grained part understanding and concise semantic and structural cues about objects, which enables many intriguing applications, such as human and animal pose estimation [6, 9, 51, 65] and robotics [73].

Over the past decades, keypoint detection has progressed significantly, from traditional image processing based methods [13, 29, 37] to deep learning based approaches [6,9,39,51,65]. The traditional methods, *e.g.*, Moravec's detector [37] and its successor Harris's detector [13], and SIFT [29], are unsupervised to detect low-level interest points but struggle to localize high-level semantic keypoints. Once we entered deep-learning era, this issue had been overcome, leading to

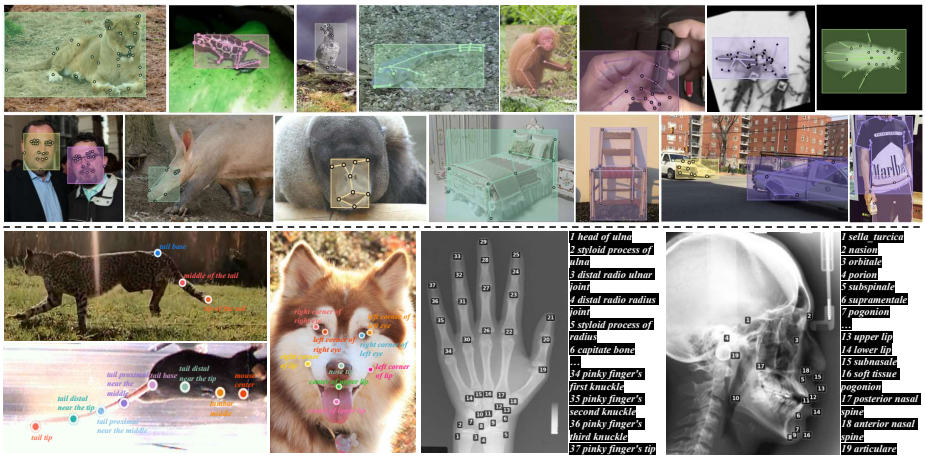

Fig. 1: A glance of MegaKPT. All images have both keypoint and text annotations. *(top)* Visualize keypoints only; *(bottom)* Visualize both keypoints and texts. The dataset covers large diversity of objects from various scenes (*best viewed in zoom*).

a series of successful convolutional neural network (CNN) or Transformer [56] based keypoint detection models, expert for various domain-specific categories, such as human pose [6,51,65] and animal pose models [69]. However, these expert models suffer from *close-set detection*, which are unable to recognize keypoints on different kinds of objects, in particular if the objects have quite different anatomies like 'fish' *vs*. 'sofa'. Therefore, an interesting question emerges: *how to build a strong general keypoint detection (GKD) model practical for diverse keypoints and objects?*

Inspired by the fact that humans can quickly recognize keypoints given one or a few visual examples, or priors with just language descriptions, recently, researchers have explored prompt based models more general than expert ones, whose goal is to detect keypoints on a query image given one or a few support images with keypoint annotations (*i.e.*, visual prompt) [15,32,33,36,64], keypoint texts (*i.e.*, text prompt) [70, 71], or both [30, 34, 66], establishing single-modal or multimodal prompted detection. However, most models are still trained in domain-specific or relatively small-scale datasets, which limits the knowledge transferability among domains and model generalization to unseen categories.

To address this issue, we firstly propose a large-scale unified keypoint dataset, dubbed ***MegaKPT***, by unifying 1.3 million diverse object instances from twenty-nine existing datasets into the same annotation format. Compared to the previous dataset UniKPT [66] which has around 418k instances, our MegaKPT not only has higher volume, but also corrects noisy annotations, supplements accurate keypoint texts and gives clear super-categories and indexes, rendering a high-quality and convenient-to-use dataset. A glance of MegaKPT is shown in Fig. 1. For the first time, we provide expert keypoint text descriptions for med-

ical scans, such as Cephalometric images for orthodontics [59] and Hand X-ray images [18], which makes zero-shot medical landmark detection easy.

To handle the large image diversity of MegaKPT and detect versatile keypoints, we further present ***GKDT***, a DINOv3 [49] based Transformer model for General Keypoint Detection. The advent of self-supervised pre-trained vision foundation model (VFM) DINOv3 [49], could serve as good weight initialization to our visual backbone to extract semantically rich dense features for keypoint localization. Inspired by ViTPose++ [65], we propose a kernel generation (KG) transformer within GKDT to convert both textual and visual keypoint representations into convolution kernels. Afterwards, these convolution kernels will be injected into detection head to perform efficient non-parametric detection. Moreover, to effectively train our model, we propose a suite of useful strategies, including i) mix-modal prompted training to enhance the consistency of prompted patterns between training and test and ii) a novel dynamic importance sampling to mitigate the data imbalance through the perspective of data sampling, which improves scores on tail classes while keeps scores on head ones.

Our contributions are in three aspects: **1)** We elaborately construct a large-scale and diverse keypoint dataset MegaKPT to support the research of general keypoint detection (GKD); **2)** Our GKDT model enjoys simplicity in architecture design, effectiveness in large-scale model training, and strong generality and performance in detecting keypoints across diverse object categories; and **3)** We propose a suite of strategies effective for GKDT model training.

To our best knowledge, this work is the first attempt of DINOv3 based model for general keypoint detection.

2 Related Work

Keypoint Detection has evolved rapidly from traditional interest point methods [13, 29] to modern deep learning-based approaches, including deep corner detection [72], semi-supervised [16,38,58], and fully supervised methods [6,9,10, 39,51,53,65]. In general, deep keypoint localization methods can be categorized into two main types: i) direct coordinate regression [7,54] and ii) heatmap-based regression with coordinate decoding [51, 65]. Unlike the existing models that are designed for recognizing specific body parts, such as top-down [51, 65] or bottom-up [6,9] human or animal pose estimators, our GKDT model is for general keypoint detection (GKD), which can deal with more versatile keypoints and objects, overcoming the limitations of close-set detection.

General Keypoint Detection is a step forward multimodal generic vision aiming at unifying the keypoint detection tasks over various object categories. To this end, the detection model must be very flexible and general. Inspired by few-shot learning [48,50,52], a series of works [32,33,36] explore visual prompted keypoint detection, whose target is to instruct the model to quickly recognize novel or base keypoints on seen or unseen objects given visual prompts. Following this line of research, Chen *et al*. [8] proposed a new weak-shot setting that transfers keyness and correspondence to novel objects via weakly labeled data.

Table 1: Statistics of **MegaKPT**. Each dataset source is credited.

Super-category	Dataset	Category	Keypoint	Image	Instance
Human pose	COCO [27]	1	17	66,808	273,469
	Human-Art [21]	1	21	50,000	123,131
Human face	300W [45]	1	68	600	600
	HELLEN [25]	1	68	2,330	2,330
	AFW [74]	1	68	337	337
	IBUG [46]	1	68	135	135
	LFPW [3]	1	68	1,035	1,035
	AFLW [23]	1	21	25,993	25,993
Human limbs	OneHand10K [60]	1	21	11,703	11,289
	HInt [41]	1	21	17,281	17,281
Animal pose	Animal [5]	5	20	4,666	6,117
	AwA pose [2]	35	39	10,064	10,064
	CUB [57]	200	15	11,788	11,788
	NABird [55]	555	11	48,562	48,562
	AP-10K [69]	54	17	10,015	13,028
	APT-36K [67]	30	17	35,708	48,704
	MacaquePose [24]	1	17	13,083	16,393
	ATRW (tiger) [26]	1	15	2,830	2,830
	AcinoSet (cheetah) [20]	1	20	5,795	5,795
	Animal Kingdom [40]	850	23	33,000	99,267
	TopViewMouse-5K [68]	1	27	5,000	5,000
Insect pose	Vinegar Fly [42]	1	32	1,500	1,500
	Desert Locust [12]	1	35	700	700
Animal face	AnimalWeb [22]	350	9	22,451	21,921
Furniture	Keypoint-5 [61]	5	8∼14	8,649	8,649
Vehicle	CarFusion [44]	3	14	53,000	100,000
Clothes	DeepFashion2 [11]	13	8∼39	491,000	491,000
Medical SC	Cephalometric [59]	1	19	400	400
	Hand X-ray [18]	1	37	910	910
Accumulated	**29**	**1587**	**740**	**935,343**	**1,348,228**

Instead of conditioning on visual prompt, the text prompted keypoint detection [70, 71] was proposed thanks to the pre-trained vision-language models (*e.g.*, CLIP [43]) which bridges the gap between vision and language. Compared to visual prompted detection, the text prompted one is more convenient and reconfigurable, saving the efforts to mark keypoints for even one example.

To embrace multimodal prompting, some multimodal models were proposed to support both image and text prompting. Lu *et al.* [34] proposed an OpenKD model, but the OpenKD was trained in small datasets and had limited generality and lower performance on large datasets as a result. Yang *et al.* [66] proposed an end-to-end X-Pose model which predicts object bounding boxes and keypoints simultaneously. However, one model may be hard to regress well for two tasks, and the performance of keypoint detection is restricted by itself object detector. Unlike X-Pose, our GKDT focuses keypoint understanding on individual object, and can couple with more advanced object detectors to deal with multi-object scenarios with higher performance.

Foundation Models represent a broad class of models pre-trained on large-scale datasets, including vision foundation models (VFMs) (*e.g.*, DINOv3 [49]), large language models (LLMs), and vision-language models (VLMs), such as CLIP [43]. Using transfer learning technologies [31, 35, 62] to adapt these models to downstream tasks in an efficient way has become popular in computer vi-

Table 2: Statistics of the unified datasets.

	MP-100	UniKPT	MegaKPT
Cite	[64]	[66]	-
Category	100	1,237	**1,587**
Keypoint	-	338	**740**
Image	20,000	226,547	**935,343**
Instance	20,000	418,487	**1,348,228**

Table 3: Generalization to unseen hand X-ray images [18] given visual, text and multimodal prompts (*i.e.*, 'both').

PCK@0.1	visual	text	both
MP-100	97.56	33.05	96.14
UniKPT	86.78	89.87	90.38
MegaKPT	**99.28**	**99.62**	**99.60**

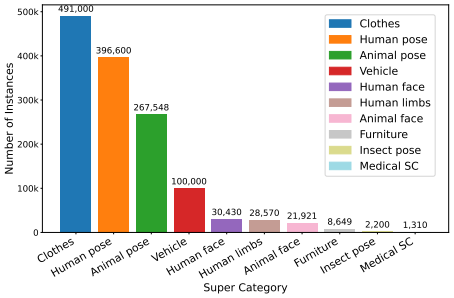

Fig. 2: Instance distribution against ten super-categories in our unified dataset.

sion [17,28,70,71]. Following this trend, we proactively explore the self-supervised DINOv3 for general keypoint detection.

3 MegaKPT: A Large-Scale High-Quality GKD Dataset

Advancing capacity and diversity of the unified keypoint dataset is crucial for general keypoint detection (GKD). Prior major datasets include MP-100 [64] and UniKPT [66]. However, MP-100 has around 200 annotated instances per object category, which makes it hard to train a robust GKD model practical for real problems. UniKPT extends MP-100 by incorporating 13 existing datasets, while the annotations are with repetitive category names, noisy indexings, unclear partition of super-categories, and cover partial instances for some datasets. These factors prevent researchers conveniently using it for model training. To address these issues and push forward the unified keypoint dataset, we build a new large-scale high-quality MegaKPT by unifying each dataset from scratch by us. The statistics are shown in Table 1.

The proposed MegaKPT has in total over 1.3 million object instances with unified annotation format across ten super-categories, with a distribution shown in Fig. 2. Each keypoint has text annotation. After filtering repetitions, MegaKPT has accumulated number of 740 unique keypoint types and 1587 object categories, spanning human and animal poses, faces and limbs, insects, furniture, vehicle, clothes, and medical scans. As shown in Table 2, MegaKPT greatly advances the prior MP-100 and UniKPT. We note that MegaKPT enjoys three features: i) large capacity of data volume; ii) large object scale variations, as small as 'vinegar fly' and 'locust' while as big as 'tiger' and 'elephant'; iii) large domain shift, with class ranging from natural scene (*e.g.*, insects and birds) to domestic scene (*e.g.*, furniture and clothes). These features would well support the GKD research and zero- or few-shot transfer to novel categories.

MegaKPT requires significant efforts to manually unify annotations into the same COCO format [27], correct ambiguous keypoint locations and texts, and supplement missing keypoint texts for those categories. Moreover, we care the symmetry and relative spatial direction of keypoints, and annotate those keypoint texts in object ego-centric view to ensure uniqueness. We also use LLM [1]

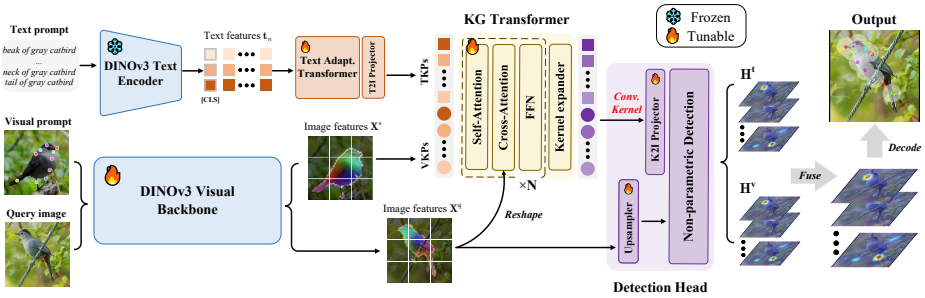

Fig. 3: Illustration of the pipeline of our GKDT model for general keypoint detection. Our model takes as input the text prompt, visual prompt, or both to detect the corresponding keypoints on a query image. Our GKDT uses a fully tuned DINOv3 vision transformer as visual backbone to extract semantically rich image features, and a kernel generation (KG) transformer to translate the textual and visual keypoint prototypes (*i.e.*, TKPs & VKPs) into convolution kernels for efficient non-parametric detection.

to standardize the object or keypoint names, *e.g.*, 'steinbucksteenbok' to 'steinbuck steenbok', and 'l_eye' to 'left eye', providing a good corpus for keypoint text. Interestingly, for the first time, we give the expert keypoint texts for medical scans of Cephalometric images for orthodontics [59] and Hand X-ray images [18], which will benefit zero-shot medical landmark detection (Fig. 1). Note that all text annotations are manually provided and examined for quality control.

To explore the dataset advantages, we reserve the entire hand X-ray images for testing while train the same GKDT model in MP-100, UniKPT, and MegaKPT, respectively. All training sets are without hand X-ray images while with real hands in daily life. Table 3 shows that the model trained in our MegaKPT dataset has highest transfer results across visual (*i.e.*, 1-shot), text (*i.e.*, 0-shot), or both prompting, validating the advancement of MegaKPT.

4 General Keypoint Detection Transformer

Following prior works [34,66], general keypoint detection (GKD) can be trained and evaluated in an episodic manner, where each episode contains a support set and a query set. The query set consists of the images to be detected, while the support set provides the prompts. When *visual prompts* (K support images with marked keypoints) are used, the task is referred to as visual-prompted keypoint detection (*i.e.*, *K-shot detection*). When only *text prompt* (language description of keypoint) is given, then the task becomes text-prompted keypoint detection (*i.e.*, *zero-shot* detection).

The overall architecture of our GKDT model is shown in Fig. 3. As one can see, our GKDT is a prompt based model, whose number of output keypoint heatmaps are corresponding to the given input prompts. As such, the model has great flexibility to process the objects that have different numbers of keypoints and anatomy structures, *e.g.*, 'tiger', 'bird', and 'sofa', *etc.*

4.1 Motivation

The generality of detecting various keypoints is fascinating. However, the question is how to extract representative keypoint prototypes from prompts to effectively guide the model in detecting keypoints in query images, thereby achieving strong performance across diverse objects. On the one hand, with the advent of vision foundation model (VFM) DINOv3 [49], it is preferable to choose it as visual backbone to extract semantically dense image features to enhance keypoint representations. As DINOv3 has been self-supervised pre-trained over a large-scale images, its weights can provide a good initialization for model learning. On the other hand, we observe that human pose expert model ViTPose++ [65] can achieve state-of-the-art results in COCO benchmark [27] even by adding a simple detection head on top of an MAE pre-trained ViT $\mathcal{F}_{\mathrm{v}}$ [14] as

$$\mathbf{H} = \mathrm{Conv}_{1\times1}(\mathcal{U}(\mathcal{F}_{\mathrm{v}}(\mathbf{I}))), \tag{1}$$

where $\mathbf{H} \in \mathbb{R}^{N\times ul\times ul}$ is the N output keypoint heatmaps corresponding to the N predefined keypoints, and $\mathcal{U}$ is an upsampler that can be bilinear interpolation or two deconv blocks to upscale the resolution of image feature map $\mathcal{F}_{\mathrm{v}}(\mathbf{I}) \in \mathbb{R}^{C\times l\times l}$ by ratio u. The $\mathbf{I} \in \mathbb{R}^{3\times l_0\times l_0}$ is the input image. By taking the convolution weights $\mathbf{W} \in \mathbb{R}^{N\times C\times1\times1}$ out from $\mathrm{Conv}_{1\times1}$ layer, we can rewrite Eq. 1 as

$$\mathbf{H} = \mathcal{U}(\mathcal{F}_{\mathrm{v}}(\mathbf{I})) \otimes \mathbf{W}. \tag{2}$$

By observing Eq. 2, one can quickly find that each filter $\mathbf{W}_i \in \mathbb{R}^{C\times1\times1}$ acts as similar role to the keypoint prototype to guide the keypoint detection. Inspired by this, if the visual and textual keypoint prototypes (*i.e.*, VKPs & TKPs) extracted from the prompts could learn keypoint representations well like convolution weights $\mathbf{W}_i$ and exhibit high similarity with image features $\mathcal{F}_{\mathrm{v}}(\mathbf{I})$, our GKDT model has potential to achieve strong performance over open-set object keypoints and categories. This idea directly motivates us to design a kernel generation (KG) transformer inside our GKDT model. In the following, we will present the working mechanism of our GKDT.

4.2 GKDT Model

As shown in Fig. 3, our GKDT mainly includes four procedures, which are i) image/text feature extraction and adaptation, ii) keypoint prototypes building, iii) kernel generation, and iv) efficient non-parametric detection.

For brevity, let us assume that the each input episode has one query image $\mathbf{I}^{\mathrm{q}}$, one support image $\mathbf{I}^{\mathrm{s}}$ with N annotated keypoints, and N keypoint texts. Firstly, we will use a DINOv3 visual backbone $\mathcal{F}_{\mathrm{v}}$ [49] to encode the support and query images as $\mathcal{F}_{\mathrm{v}}(\mathbf{I}^{\mathrm{s}})$ and $\mathcal{F}_{\mathrm{v}}(\mathbf{I}^{\mathrm{q}})$ in deep feature space $\mathbb{R}^{l\times l\times d}$. Moreover, with the DINOv3 text encoder trained by dino.txt $\mathcal{F}_{\mathrm{t}}$ [19], the N texts are firstly tokenized, and then encoded as text features $\mathbf{t}_n \in \mathbb{R}^{m\times d}$, where $n = 1, 2, \cdots, N$; and m is sequence length. For the whole DINOv3 visual backbone, we *fully tune* it to unleash the power of self-supervised learned visual knowledge inside

the model for keypoint localization. However, considering the total capacity of keypoint texts are relatively small compared to LLM corpus, we freeze the entire DINOv3 text encoder, and only finetune $\mathbf{t}_n$ with an additional text adaptation transformer $\mathcal{A}_\text{t}$ as $\mathbf{t}_n := \mathcal{A}_\text{t}(\mathbf{t}_n)$. There is no additional visual adaptation module as the entire DINOv3 visual backbone is tuned. Since only DINOv3 ViT-L has the paired text encoder, to make our model be flexible to accommodate various DINOv3 visual backbone variants, we insert a T2I projector that always maps the text features to have the same channel dimension with the image features, thus handling the feature channel inconsistency.

Afterwards, we extract textual keypoint representations by taking the classification token $\mathbf{\Phi}_n^\text{t} \in \mathbb{R}^d$ from each text feature $\mathbf{t}_n \in \mathbb{R}^{m\times d}$. Then, we perform average to obtain textual keypoint prototype (TKP) as $\mathbf{\Psi}_n^\text{t}$ if multiple texts are given. For visual keypoint representations (VKR), we encode each support keypoint $\mathbf{p}_n$ into a Gaussian heatmap $\mathbf{H}(\mathbf{p}_n;\sigma)$ and obtain its visual embedding $\mathbf{\Phi}_n^\text{v} \in \mathbb{R}^d$ via linear-weighted summation between $\mathbf{H}(\mathbf{p}_n;\sigma)$ and $\mathcal{F}_\text{v}(\mathbf{I}^\text{q})$. Here, σ is the standard deviation that controls the spread of the Gaussian. If K support images are provided, the VKRs of the same keypoint type, denoted as $\mathbf{\Phi}_{k,n}^\text{v}$, are averaged to produce the visual keypoint prototype (VKP) as $\mathbf{\Psi}_n^\text{v} = \frac{1}{K}\sum_k \mathbf{\Phi}_{k,n}^\text{v}$.

Motivated by Section 4.1, all prototypes are combined as a set of tokens and fed into a KG transformer $\mathcal{K}$ for feature refinement and kernel generation. The $\mathcal{K}$ includes transformers followed by a kernel expander. The self-attention in transformer allows TKPs and VKPs to exchange information and learn relations from each other. In this way, the strong modal prompt, either visual or text prompt, can help the other weak one and improve multimodal prompted performance. To further enhance keypoint representations and improve correlation between the prompt and query image features, we add cross-attention to aggregate more context information from query image. After layer-wise deconv via a kernel expander, the KG will output a set of convolution kernels $\{\mathbf{W}_n^\text{v}\} \cup \{\mathbf{W}_n^\text{t}\} = \mathcal{K}(\{\mathbf{\Psi}_n^\text{v}\} \cup \{\mathbf{\Psi}_n^\text{t}\}, \mathcal{F}_\text{v}(\mathbf{I}^\text{q}))$, where $\mathbf{W}_n \in \mathbb{R}^{C\times s\times s}$ is with kernel size s. Then, the kernels are projected to the same channel with query image feature via a K2I projector, and performs efficient non-parametric detection as:

$$\begin{aligned} \mathbf{H}^\text{v} &= \text{norm}(\mathcal{U}(\mathcal{F}_\text{v}(\mathbf{I}^\text{q}))) \otimes \text{norm}(\mathbf{W}^\text{v})/s^2 \\ \mathbf{H}^\text{t} &= \text{norm}(\mathcal{U}(\mathcal{F}_\text{t}(\mathbf{I}^\text{q}))) \otimes \text{norm}(\mathbf{W}^\text{t})/s^2 \end{aligned}, \tag{3}$$

where $\text{norm}(\cdot)$ refers to l_2 normalize in channel dimension. In testing, both the visual or textual prompt induced heatmaps $\mathbf{H}^\text{v}$ and $\mathbf{H}^\text{t}$ are fused to yield an ensemble heatmap output as $\mathbf{H} = (\mathbb{I}^\text{v} \odot \mathbf{H}^\text{v} + \mathbb{I}^\text{t} \odot \mathbf{H}^\text{t})/(\mathbb{I}^\text{v} + \mathbb{I}^\text{t})$, where $\mathbb{I}^\text{v}$ (or $\mathbb{I}^\text{t}$) is the validity indicator of keypoints, whose entry $\mathbb{I}_n^\text{v}$ (or $\mathbb{I}_n^\text{t}$) is 1 if the union of n-th keypoint in visual prompt (or text prompt) is valid, otherwise is 0. In training, for induced heatmaps, we will construct GT heatmaps $\mathbf{H}^*$ for supervision, resulting in the loss $\mathcal{L} = \mathbb{E}\big(\|\mathbb{I}^\text{v} \odot (\mathbf{H}^\text{v} - \mathbf{H}^*)\|_\text{F}^2 + \|\mathbb{I}^\text{t} \odot (\mathbf{H}^\text{t} - \mathbf{H}^*)\|_\text{F}^2\big)/(\mathbf{1}^\intercal\mathbb{I}^\text{v} + \mathbf{1}^\intercal\mathbb{I}^\text{t})$, where $\mathbf{1}$ is all-one vector and $\mathbf{H}_n^* = \mathbf{H}(\mathbf{p}_n^*;\sigma)$ if the corresponding query keypoint $\mathbf{p}_n^*$ exists; otherwise $\mathbf{H}_n^* = \mathbf{0}$ which forms negative keypoints learning.

4.3 Strategies for Model Training

To effectively train our model, we propose a suite of strategies as follows:

1) Mix-modal prompted training: When applying the model in real-world scenarios, users may perform visual, text, or both prompting given their preferences. If one only adopts one mode of prompt during training, *e.g.* 'both' one, it will inevitably cause prompt inconsistency between training and test. To mimic real prompted behaviors, we propose mix-modal prompted training from the perspective of prompt sampling, where each training episode randomly selects a mode in set {`visual`, `text`, `both`} (default set), and then mask keypoint indicators $\mathbb{I}^{\text{v}}$ and $\mathbb{I}^{\text{t}}$ accordingly. We also studied prompt sets such as {`visual`, `text`} and {`both`}, and empirically found the default set gives best overall scores.

2) Dynamic importance sampling: Fig. 2 shows that our MegaKPT exhibits long-tail phenomenon, where the head classes take dominant number of instances while the tail ones take minor. Such a data imbalance widely exists and presents a challenge for general model training. A common way to address this issue is to perform data balancing. However, it will lose many precious data samples for training and weakens model capability. A nature question is if we could directly train models on such an imbalanced dataset, while not only keep strong performance on head classes but also improve scores on tail ones.

We try to answer this question through the perspective of data sampling. Consider data sampling without replacement in a data pool $\mathcal{D} = \{D_i\}_{i=1}^{S}$, where $\mathcal{D}_i$ is the data partition of i-th super-category. If one uses uniform sampling, the data samples in tail classes will be quickly exhausted, risking the model in catastrophic forgetting tail classes at late training stage. If using importance sampling, the samples are mostly from head classes, leading to underfitting in tail ones. To reach a balance, we propose a novel dynamic importance sampling, whose key is to dynamically judge if the super-category c of the episode $\mathbf{e}$ sampled via importance $p_c = |\mathcal{D}_c|/\mathcal{D}$ comes from head classes, *i.e.*, those classes with top-γS number of instances, where $\gamma = 0.5$ is a fix ratio. If it is, the drawn samples $\mathbf{e}$ will be removed from $\mathcal{D}_c$; otherwise, it will be placed back. If $\mathcal{D}_c$ is empty, the super-category c will be also removed with an update $\mathcal{D} := \mathcal{D} \backslash \mathcal{D}_c$ and $S := S - 1$. Note that head categories will be dynamically changed based on their remaining number of instances. After iteratively sampling, the whole distribution is gradually balanced. We visualize the distributions in simulated sampling and provide detailed algorithm in **§A of Suppl.**

5 Experiments

5.1 Experiment Setup

Dataset Splits: For fair comparison, we follow original partitions for those datasets in MegaKPT that have official training, validation and test splits, such as COCO [27] and Human-Art [21]. Otherwise, we perform splits as follows:

- Following previous work [34], the AwA pose dataset [2] is split into 25 species for training and the remaining 10 disjoint ones for testing; The bird datasets

Table 4: General keypoint detection across thirteen datasets in single-object scenario. The PCK@0.1 score is reported. The test sets that are unseen (▲) and generalized unseen (△) are marked. Symbol * means results reproduced via official released model.

Model	Prompt	Animal (△)	AwA (△)	CUB (▲)	NABird (▲)	AP-10K (△)	Vinegar fly	Locust	Mouse5k	Macaque	Tiger	Animal Kin.	COCO val	HumanArt
DINOv3 [49]	visual	58.24	60.55	80.73	71.11	59.74	80.19	78.10	50.49	43.32	54.17	40.63	39.15	43.34
CapeF. [47]	visual	41.48	51.65	83.91	87.91	67.00	85.36	80.28	80.08	61.32	60.13	65.61	66.78	62.71
OpenKD [34]	visual	49.41	61.54	85.51	86.92	61.92	90.87	90.82	85.10	66.60	69.66	58.71	65.26	60.06
GKDT (Ours)	visual	**81.11**	**84.64**	**97.71**	**96.25**	**89.25**	**97.59**	**98.66**	**96.40**	**88.82**	**92.33**	**90.74**	**88.64**	**82.92**
DINOv3 [49]	text	20.72	13.82	10.60	8.57	21.01	3.24	2.99	5.29	17.45	22.62	5.52	3.13	2.19
OpenKD [34]	text	55.86	81.80	89.82	91.86	75.39	96.95	97.77	92.76	85.40	87.55	78.61	82.40	77.01
X-Pose* [66]	text	40.69	28.57	67.07	39.62	77.23	98.13	98.59	16.01	91.60	55.10	54.75	86.05	73.06
GKDT (Ours)	text	**73.27**	**92.80**	**98.58**	**96.94**	**91.93**	**98.91**	**99.49**	**97.37**	**93.75**	**95.88**	**93.77**	**93.59**	**90.48**
DINOv3 [49]	both	35.80	30.77	44.37	31.40	38.64	23.18	21.52	20.61	33.58	41.18	14.91	9.36	7.36
OpenKD [34]	both	57.58	80.37	90.82	91.81	74.65	95.97	96.69	92.31	83.89	86.18	77.54	81.35	75.62
GKDT (Ours)	both	**76.87**	**91.14**	**98.48**	**96.86**	**91.50**	**98.79**	**99.37**	**97.29**	**93.15**	**95.40**	**93.29**	**93.52**	**89.68**

Table 5: General keypoint detection in single-object scenario in nine more datasets.

Model	Prompt	Faces & hands				Furni., vehi., clothes			Medical	
		300W	Ani.Web (▲)	OneHand	HInt	Keypoint-5	CarFusion	D.Fashion2	Cephalo	Hand X-ray (▲)
DINOv3 [49]	visual	65.01	47.37	29.31	15.57	44.89	32.19	41.28	81.67	91.68
CapeF. [47]	visual	89.47	71.83	55.33	36.38	53.21	74.83	78.76	67.40	47.66
OpenKD [34]	visual	86.98	64.47	51.43	25.66	67.17	81.10	73.94	97.89	68.67
GKDT (Ours)	visual	**96.99**	**82.54**	**92.56**	**69.79**	**83.01**	**94.15**	**90.96**	**99.46**	**99.28**
DINOv3 [49]	text	16.41	8.97	7.01	4.93	2.66	0.27	1.92	0.39	1.19
OpenKD [34]	text	96.63	73.34	75.70	52.26	80.32	91.85	87.30	99.22	82.24
X-Pose* [66]	text	**99.53**	27.35	73.05	40.42	46.02	87.81	66.84	3.54	2.41
GKDT (Ours)	text	99.04	**86.36**	**95.36**	**82.64**	**88.15**	**96.25**	**95.34**	**99.49**	**99.62**
DINOv3 [49]	both	36.73	22.11	14.11	9.69	15.74	2.79	8.00	16.08	25.95
OpenKD [34]	both	95.67	73.10	73.88	50.94	77.87	90.58	85.82	99.18	85.31
GKDT (Ours)	both	**98.87**	**86.24**	**95.12**	**81.13**	**87.04**	**95.92**	**94.58**	**99.49**	**99.60**

CUB [57] and NABird [55] are split as 100, 50, 50 and 333, 111, 111 for training, validation, and test, respectively.
- The AP-10K [69] uses 42 animal species for training and the remaining 12 species for test. The animal face dataset AnimalWeb [22] is split into 250 categories as seen classes and the other 100 as unseen ones for test.
- For the purpose of testing model generalization, the entire Animal pose dataset [5] and Hand X-ray [18] are reserved for testing.

The full splits are in **§B of Suppl.**. Note that test sets of CUB, NABird, AnimalWeb and Hand X-ray are *unseen* to training data. We mark them with (▲) for distinctions. Moreover, the test sets of Animal pose, AwA and AP-10K are marked with (△), as they are *generalized unseen*[1] with possible overlapping to training data. The training sets are combined for GKD model training.

Metrics: Both percentage of correct keypoints (PCK) [32] and average precision (AP) [27] (*e.g.*, for COCO and Human-Art) are used for rigorous evaluation.

[1] 'Generalized unseen' refers that a data sample is from seen or unseen classes [63].

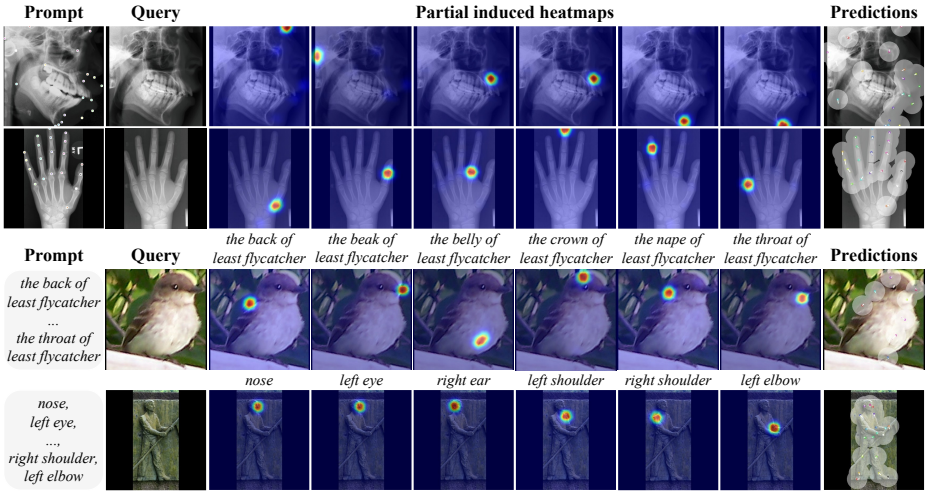

Fig. 4: Visualization of visual and text prompted keypoint detection using our GKDT model. The white shadow in predictions signifies PCK@0.1, which means a predicted keypoint (*i.e.*, circle) is correct if falling on this area. GT are marked as tilted crosses.

Compared Methods: The compared methods include those that can perform general keypoint detection (GKD) such as the training-free vanilla DINOv3 [49], OpenKD [34], and X-Pose [66], and the visual prompt based method CapeFormer [47] and the methods that are expert for human pose estimation such as state-of-the-art model ViTPose++-H [65]. For OpenKD and CapeFormer, we adopt their source codes to fairly train and test in MegaKPT. X-Pose has no training codes as well as visual branches in released model, thus we only compare it for text prompted detection. With different DINOv3 visual backbones, *e.g.*, DINOv3-S/B/L/H, our GKDT has variants GKDT-S/B/L/H accordingly. We use GKDT-L as our default model (*i.e.*, *GKDT*).

Implementation Details: The size of input images to GDKT models is 384×384. Our KG module $\mathcal{K}$ uses two transformer blocks and text adaptation $\mathcal{A}_{\text{t}}$ uses one. The size of generated kernels s performs well at 1×1 and 3×3. All our GKDT models are trained by 20 epochs using 16 H800 GPUs.

5.2 General Keypoint Detection in Single-Object Scenario

We first perform extensive experiments on general keypoint detection in 22 test sets (Table 4 & Table 5), where each input image has a single object cropped via GT bounding box. 1000 episodes with each 10 query images are tested. We use PCK@0.1 as a unified metric for convenient evaluation for all test sets. For COCO and HumanArt, we also report AP [27] scores in Sec. 5.3. Table 4 shows that our GKDT model consistently achieves best scores and outperforms vanilla DINOv3, OpenKD, and X-Pose under whether visual, text, or both prompts across 13 datasets, covering the super-categories of animal, insect, and human

poses. Moreover, we evaluates 9 more datasets covering human and animal faces and hands, furniture, vehicle, clothes, and medical images in Table 5. Again, our GKDT shows strong results using a single model. Despite that some test sets are unseen (▲) or generalized unseen (△), our model shows good generalization. For instance, our GKDT achieves 98.48% in CUB and 96.86% in NABird in localizing the landmarks on unseen birds, 86.24% in AnimalWeb in unseen animal faces, and 99.60% for hand X-ray images under multimodal prompting (*i.e.*, 'both' one). We suspect that the reason behind the surprising scores on hand X-ray is because medical images are captured in structured scene thus being easier compared to those real hand poses from complex daily scenarios. Together, 17 out of 22 datasets have over 90% accuracy under multimodal prompting, showing that our GKDT model has strong generality and practical applicability in a broad set of categories. Fig. 4 gives a glance of visualizations.

5.3 General Keypoint Detection in Multi-Object Scenario

Our GKDT models focus on keypoint detection for individual objects. However, it can be also applied to multi-object scenario via two-stage top-down method, which firstly localizes object bounding boxes via an object detector and then detects the keypoints for each box. Compared to end-to-end method, two-stage one can couple with more advanced object detectors for high-quality detection.
Results on multi-human pose estimation: Firstly, we evaluate our GKDT models in zero-shot detection in COCO [27] and HumanArt [21] validation sets. By coupling with open-set object detector Grounding DINO [28] (G-DINO), our GKDT and GKDT-H yields 75.0% and 77.2% in COCO (Table 6a), setting the state-of-the-art results in general models. In COCO object detection, Faster RCNN, G-DINO, and GT has AP of 56.4%, 65.4%, and 100.0%. By using better bounding boxs, our GKDT strikes higher results. After finetuning, our GKDT-H improves scores further (78.1% *vs*. 77.2%). In HumanArt, our GKDT and GKDT-H with G-DINO obtain scores of 70.0% and 71.3%, while with GT

Table 6: Multi-human pose estimation on COCO and Human-Art val sets. General models use text prompts. G-DINO: Grounding DINO [28]; *: results reproduced via official released model; †: finetuned on COCO+HumanArt; l. param: learnable params.

Models	Box Detector	AP	AP^{50}	AP^{75}	l. param
Expert models					
ViTPose-H [65]	Faster RCNN	79.1	91.6	85.7	637.2M
General models					
X-Pose* [66]	End-to-end	71.8	88.9	78.3	128.1M
GKDT (Ours)	Faster RCNN	73.2	88.7	80.1	349.1M
GKDT (Ours)	G-DINO	75.0	90.8	82.3	349.1M
GKDT (Ours)	GT box	76.5	93.7	83.9	349.1M
GKDT-H (Ours)	G-DINO	77.2	91.6	84.3	898.4M
GKDT-H (Ours)	GT box	78.2	94.5	84.7	898.4M
GKDT-H† (Ours)	G-DINO	78.1	91.7	85.1	898.4M

(a)

Models	Box Detector	AP
Expert models		
ViTPose-H [65]	GT	67.7
General models		
X-Pose* [66]	End-to-end	70.7
GKDT (Ours)	G-DINO	70.0
GKDT (Ours)	X-Pose	73.4
GKDT (Ours)	GT box	78.1
GKDT-H (Ours)	G-DINO	71.3
GKDT-H (Ours)	GT box	79.9
GKDT-H† (Ours)	G-DINO	72.0

(b)

Table 7: Results on four multi-object datasets. The scores of PCK@{0.05, 0.10, 0.20} are reported. Object boxes are filtered by the score threshold that gives best F-measure.

Model	Box Det.	AP-10K			Macaque Pose			Mouse5K			Carfusion		
		0.05	0.10	0.20	0.05	0.10	0.20	0.05	0.10	0.20	0.05	0.10	0.20
X-Pose* [66]	End-to-end	66.29	76.06	82.36	94.11	97.41	98.70	24.76	35.43	44.57	**98.96**	**99.51**	99.78
GKDT (Ours)	G-DINO	**85.59**	**92.60**	**95.56**	**94.66**	**97.57**	**98.78**	**80.05**	**82.89**	**87.37**	98.71	99.47	**99.94**

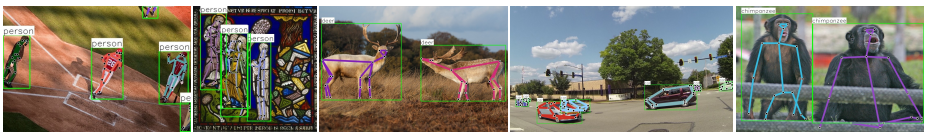

Fig. 5: Visualization of general keypoint detection in multi-object scenario via GKDT.

boxes have 78.1% and 79.9%, showing significant room to improve. The G-DINO performs modest in HumanArt. After using the identical boxes from X-Pose, our GKDT improves the results compared to X-Pose (73.4% *vs*. 70.7%).

Results on animals and vehicles: We evaluate on more multi-object datasets with PCK under different thresholds. The lower the threshold, the stricter the localization precision. Table 7 shows that our GKDT model performs quite well, in particular outperforms X-Pose in AP-10K and TopviewMouse5K by a large margin. Note that all the results are achieved via a single model. Fig. 5 shows the examples of applying our GKDT in multi-object scenarios.

5.4 Ablation Study

Study on pre-trained VFMs: Besides DINOv3, we investigate using other pre-trained VFMs as visual backbone such as CLIP [4] and MAE ViTs [14]. Both DINOv3 and MAE ViTs are self-supervised, while CLIP is weakly-supervised by languages. Fig. 6a shows that DINOv3 models give best GKD performance and DINOv3-L has performance close to DINOv3-H.

Mix-modal prompted training: Fig. 6b shows if only using prompt set {`both`} during training, the single-modal prompted testing, either visual or text, is suboptimal. If using {`visual`, `text`}, the testing with both prompt drops. By sampling from the set {`visual`, `text`, `both`} yields best consistency and results.

Data amounts & multimodal prompting: The self-supervised learned visual knowledge and priors of DINOv3 enable strong transfer to GKD. Fig. 6c shows that even using 10% of training data in Cephalometric dataset, our GKDT obtains over 90% accuracy. Moreover, in real-world testing, one may not know which kind of prompt is optimal. The advantage of using multimodal prompt (*i.e.*, 'both' one) can mitigate the risk of choosing a weak-modal prompt and strongly maintain performance compared to the strongest modal prompt, either visual or text ones. In evaluation of using multimodal prompt, Table 4 and Table 5 show that 21 out of 22 tasks have $< 2\%$ drops compared to the strongest modal prompt, and 18 out of 22 tasks have $< 1\%$ drop.

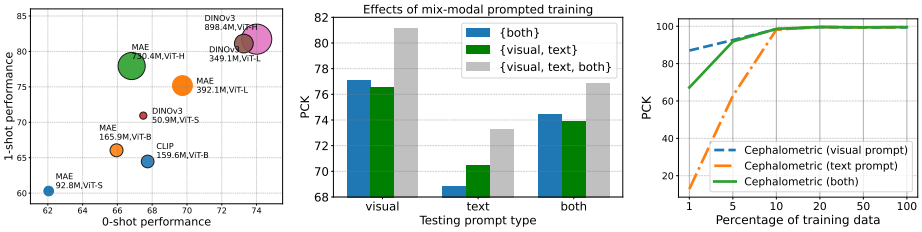

Fig. 6: Study on pre-trained VFMs (a), mix-modal prompted training (b), and amounts of training data (c). Animal pose dataset is tested in (a) and (b). PCK@0.1 is used.

Table 8: Ablations on model components. (a) Impacts of tuning $\mathcal{F}_\text{v}$ and using KG transformer $\mathcal{K}$; (b) Context aggregation in $\mathcal{K}$. SA: self-attention; CA: cross-attention.

Model		Animal pose		CUB		HInt	
Tune $\mathcal{F}_\text{v}$	KG	visual	text	visual	text	visual	text
		53.86	54.00	79.55	87.11	12.15	40.48
	✓	64.22	63.88	77.39	97.23	19.96	58.57
✓		79.08	68.29	96.78	96.81	64.74	82.83
✓	✓	81.11	73.27	97.71	98.58	69.79	82.64

(a)

KG		HInt	
SA	CA	visual	text
		64.74	82.83
✓		69.05 (+4.31)	82.15
	✓	65.90 (+1.16)	82.94
✓	✓	69.79 (+5.05)	82.64

(b)

Study on model components: Table 8a shows that if without tuning visual backbone $\mathcal{F}_\text{v}$ and using KG transformer $\mathcal{K}$, the results are worst. After tuning $\mathcal{F}_\text{v}$, the scores boost greatly. In both cases of w. and w/o tuning $\mathcal{F}_\text{v}$, KG transformer enhances scores and mitigates failures. The self-attention (SA) in $\mathcal{K}$ learns relations and aggregates context among visual and textual keypoint representations, while cross-attention (CA) in $\mathcal{K}$ aggregates context from query image features. To understand their roles, we ablate them. Table 8b shows both SA and CA contribute positive impacts, but SA is dominant. Interestingly, the improved scores are in visual prompting, which shows the strong modal prompt (*i.e.*, text one in this case) could improve the weak one.

Dynamic importance sampling: Table 9a shows that our dynamic importance sampling greatly enhances model performance in tail classes while keeps scores in head ones, boosting OpenKD by 3.92% and 6.36% in multimodal prompting in HInt and Cephalometric datasets, respectively, and in our GKDT by 1.03% and 0.40%. We notice that the benefits brought to GKDT are less than OpenKD. We conjecture the reason is as GKDT model is much stronger, thus improvement room is smaller and harder than OpenKD.

Cross-supercategory transfer: Table 9b shows our GKDT model transfers well from human to macaque as they share similar structures (92.82% *vs*. 81.03%), and as expected, having lower gap compared to four-footed animals in AP-10K.

Table 9: Study on data sampling strategies and cross-supercategory transfer. (a) Comparison of uniform, importance, and our dynamic importance sampling; (b) Transfer between Human Pose (HP) and Animal Pose (AP) supercategories. S: Source domain; T: Target domain. PCK@0.1 scores under 'both' prompt are reported.

Model	Strategy	Head		Tail	
		DF2	COCO	HInt	Cephalo
OpenKD	Uniform	85.23	81.60	47.02	90.99
	Importance	85.44	81.72	46.64	92.82
	Dynamic	85.82	81.35	50.94 (+3.92)	99.18 (+6.36)
GKDT	Uniform	94.53	93.57	78.18	99.09
	Importance	94.70	93.40	80.10	98.43
	Dynamic	94.58	93.52	81.13 (+1.03)	99.49 (+0.40)

(a)

HP to AP	PCK
COCO (S)	92.82
Macaque (T)	81.03
AP-10K (T)	58.87
AP to HP	**PCK**
Macaque (S)	90.30
AP-10K (S)	90.41
COCO (T)	82.31

(b)

6 Conclusion

We advance the unified keypoint dataset to a new level, proposing the MegaKPT to support the research on general keypoint detection. Moreover, our developed DINOv3 based general keypoint detection model, called GKDT, achieves state-of-the-art results in most single-object and multi-object scenarios under whether visual, text, or multimodal prompting, with the help of a suite of useful strategies for model training. We believe that our work will pave the way for GKD research and offers great practicality in real-world applications.

Acknowledgement

The authors would like to thank Martin Urschler and Simon Johannes Joham for providing the Hand X-ray dataset, and the helpful suggestions from anonymous reviewers. This work was supported in part by the Research Grants Council under the Areas of Excellence scheme grant AoE/E-601/22-R.

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. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. IEEE transactions on pattern analysis and machine intelligence **35**(12), 2930–2940 (2013)
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. Chen, J., Luo, Z., Liu, Z., Jiang, W., Niu, L., Fang, Y.: Weak-shot keypoint estimation via keyness and correspondence transfer. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
9. 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)
10. 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)
11. Ge, Y., Zhang, R., Wang, X., Tang, X., Luo, P.: Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5337–5345 (2019)
12. Graving, J.M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B.R., Couzin, I.D.: Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. elife **8**, e47994 (2019)
13. Harris, C., Stephens, M., et al.: A combined corner and edge detector. In: Alvey vision conference. vol. 15, pp. 10–5244. Manchester, UK (1988)
14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
15. Hirschorn, O., Avidan, S.: A graph-based approach for category-agnostic pose estimation. In: European Conference on Computer Vision. pp. 469–485. Springer (2024)
16. 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)
17. 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)
18. Joham, S.J., Hadzic, A., Urschler, M.: Implicit is not enough: Explicitly enforcing anatomical priors inside landmark localization models. Bioengineering **11**(9), 932 (2024)
19. Jose, C., Moutakanni, T., Kang, D., Baldassarre, F., Darcet, T., Xu, H., Li, D., Szafraniec, M., Ramamonjisoa, M., Oquab, M., et al.: Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24905–24916 (2025)
20. Joska, D., Clark, L., Muramatsu, N., Jericevich, R., Nicolls, F., Mathis, A., Mathis, M.W., Patel, A.: Acinoset: a 3d pose estimation dataset and baseline models for cheetahs in the wild. In: 2021 IEEE international conference on robotics and automation (ICRA). pp. 13901–13908. IEEE (2021)

21. Ju, X., Zeng, A., Wang, J., Xu, Q., Zhang, L.: Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 618–629 (2023)
22. Khan, M.H., McDonagh, J., Khan, S., Shahabuddin, M., Arora, A., Khan, F.S., Shao, L., Tzimiropoulos, G.: Animalweb: A large-scale hierarchical dataset of annotated animal faces. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6939–6948 (2020)
23. Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: 2011 IEEE international conference on computer vision workshops (ICCV workshops). pp. 2144–2151. IEEE (2011)
24. Labuguen, R., Matsumoto, J., Negrete, S.B., Nishimaru, H., Nishijo, H., Takada, M., Go, Y., Inoue, K.i., Shibata, T.: Macaquepose: a novel "in the wild" macaque monkey pose dataset for markerless motion capture. Frontiers in behavioral neuroscience **14**, 581154 (2021)
25. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: European conference on computer vision. pp. 679–692. Springer (2012)
26. Li, S., Li, J., Tang, H., Qian, R., Lin, W.: Atrw: a benchmark for amur tiger re-identification in the wild. arXiv preprint arXiv:1906.05586 (2019)
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
28. 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)
29. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)
30. Lu, C.: General Keypoint Detection: Few-Shot and Zero-Shot. Ph.D. thesis, The Australian National University (Australia) (2024)
31. 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)
32. 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)
33. 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)
34. Lu, C., Liu, Z., Koniusz, P.: Openkd: Opening prompt diversity for zero-and few-shot keypoint detection. In: European Conference on Computer Vision. pp. 148–165. Springer (2024)
35. Lu, C., Wang, H., Gu, C., Wu, K., Guan, X.: Viewpoint estimation for workpieces with deep transfer learning from cold to hot. In: International Conference on Neural Information Processing. pp. 21–32. Springer (2018)
36. Lu, C., Zhu, H., Koniusz, P.: Exploiting class-agnostic visual prior for few-shot keypoint detection. International Journal of Computer Vision **134**(2), 63 (2026)
37. Moravec, H.P.: Obstacle avoidance and navigation in the real world by a seeing robot rover. Stanford University (1980)

38. Moskvyak, O., Maire, F., Dayoub, F., Baktashmotlagh, M.: Semi-supervised keypoint localization. arXiv preprint arXiv:2101.07988 (2021)
39. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016)
40. Ng, X.L., Ong, K.E., Zheng, Q., Ni, Y., Yeo, S.Y., Liu, J.: Animal kingdom: A large and diverse dataset for animal behavior understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19023–19034 (2022)
41. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3d with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9826–9836 (2024)
42. 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)
43. 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)
44. Reddy, N.D., Vo, M., Narasimhan, S.G.: Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicles. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1906–1915 (2018)
45. Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: Database and results. Image and vision computing **47**, 3–18 (2016)
46. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 397–403 (2013)
47. Shi, M., Huang, Z., Ma, X., Hu, X., Cao, Z.: Matching is not enough: A two-stage framework for category-agnostic pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7308–7317 (2023)
48. Shi, W., Lu, C., Shao, M., Zhang, Y., Xia, S., Koniusz, P.: Few-shot shape recognition by learning deep shape-aware features. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1848–1859 (2024)
49. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
50. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. arXiv preprint arXiv:1703.05175 (2017)
51. 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)
52. 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/CVF Conference on Computer Vision and Pattern Recognition. pp. 1199–1208 (2018)
53. 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)
54. 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)

55. 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)
56. 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)
57. 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)
58. 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)
59. Wang, C.W., Huang, C.T., Lee, J.H., Li, C.H., Chang, S.W., Siao, M.J., Lai, T.M., Ibragimov, B., Vrtovec, T., Ronneberger, O., et al.: A benchmark for comparison of dental radiography analysis algorithms. Medical image analysis **31**, 63–76 (2016)
60. Wang, Y., Peng, C., Liu, Y.: Mask-pose cascaded cnn for 2d hand pose estimation from single color image. IEEE Transactions on Circuits and Systems for Video Technology **29**(11), 3258–3268 (2018)
61. Wu, J., Xue, T., Lim, J.J., Tian, Y., Tenenbaum, J.B., Torralba, A., Freeman, W.T.: Single image 3d interpreter network. In: European Conference on Computer Vision. pp. 365–382. Springer (2016)
62. Wu, X., Lu, C., Gu, C., Wu, K., Zhu, S.: Domain adaptation for viewpoint estimation with image generation. In: 2021 International Conference on control, automation and information sciences (ICCAIS). pp. 341–346. IEEE (2021)
63. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning-the good, the bad and the ugly. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4582–4591 (2017)
64. 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)
65. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose++: Vision transformer for generic body pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(2), 1212–1230 (2023)
66. Yang, J., Zeng, A., Zhang, R., Zhang, L.: X-pose: Detecting any keypoints. In: European Conference on Computer Vision. pp. 249–268. Springer (2024)
67. Yang, Y., Yang, J., Xu, Y., Zhang, J., Lan, L., Tao, D.: Apt-36k: A large-scale benchmark for animal pose estimation and tracking. Advances in Neural Information Processing Systems **35**, 17301–17313 (2022)
68. Ye, S., Filippova, A., Lauer, J., Vidal, M., Schneider, S., Qiu, T., Mathis, A., Mathis, M.W.: Superanimal models pretrained for plug-and-play analysis of animal behavior. arXiv preprint arXiv:2203.07436 **4**(5) (2022)
69. Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., Tao, D.: Ap-10k: A benchmark for animal pose estimation in the wild. arXiv preprint arXiv:2108.12617 (2021)
70. Zhang, H., Xu, L., Lai, S., Shao, W., Zheng, N., Luo, P., Qiao, Y., Zhang, K.: Open-vocabulary animal keypoint detection with semantic-feature matching. International Journal of Computer Vision **132**(12), 5741–5758 (2024)
71. 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)

72. Zhao, S., Gong, M., Zhao, H., Zhang, J., Tao, D.: Deep corner. International Journal of Computer Vision pp. 1–25 (2023)
73. Zhou, B., Zhou, H., Liang, T., Yu, Q., Zhao, S., Zeng, Y., Lv, J., Luo, S., Wang, Q., Yu, X., et al.: Clothesnet: An information-rich 3d garment model repository with simulated clothes environment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20428–20438 (2023)
74. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 2879–2886. IEEE (2012)