Generalizable Object Keypoint Localization from Generative Priors

Dongkai Wang[1,2] Jiang Duan[1] Liangjian Wen[1] Shiyu Xuan[3] Hao Chen[4] Shiliang Zhang[4,5]

[1]Southwestern University of Finance and Economics [2]Engineering Research Center of Intelligent Finance, Ministry of Education

[3]School of Computer Science and Engineering, Nanjing University of Science and Technology [4]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University [5]Peng Cheng Laboratory, Shenzhen, China

{wdk, duanj_t, wenlj}@swufe.edu.cn, shiyu_xuan@njust.edu.cn, {hchen, slzhang.jdl}@pku.edu.cn

Abstract

Generalizable object keypoint localization is a fundamental computer vision task in understanding the object structure. It is challenging for existing keypoint localization methods because their limited training data cannot provide generalizable shape and semantic cues, leading to inferior performance and generalization capability. Instead of relying on large scale training data, this work tackles this challenge by exploiting the rich priors from large generative models. We propose a data-efficient generalizable localization method named GenLoc. GenLoc extracts the generative priors from a pre-trained image generation model by calculating the correlation map between image latent feature and condition embedding. Those priors are hence optimized with our proposed heatmap expectation loss to perform object keypoint localization. Benefited by the rich knowledge of generative priors in understanding of object semantics and structures, GenLoc achieves superior performance on various object keypoint localization benchmarks. It shows more substantial performance enhancements in cross-domain, few-shot and zero-shot evaluation settings, e.g., getting 20%+ AP enhancement over CLAMP [43] in various zero-shot settings.

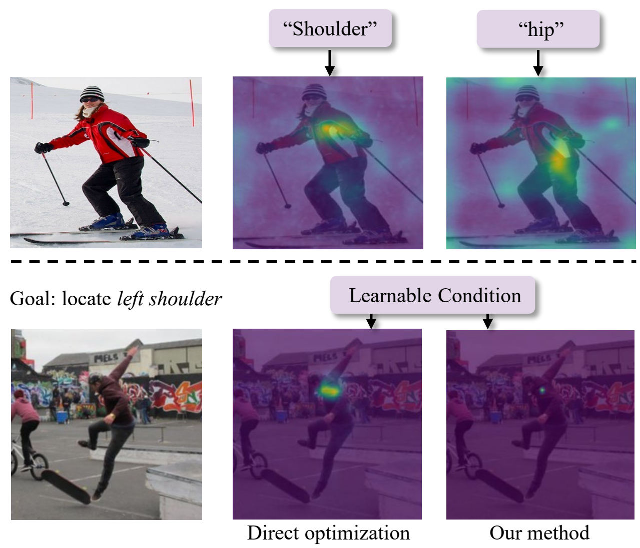

Figure 1. Upper: The pre-trained image generation model LDM [25] has shown the correlation between input image latent feature and different text condition through the cross attention mechanism. Bottom: We aim to optimize the correlation map to locate the object keypoint by learning a specific condition embedding. Directly optimizing the cross attention map in LDM cannot produce well-localized heatmap. Our method can address this issue and produce more accurate results.

1. Introduction

The rapid progress in deep learning and computer vision has pushed the object visual perception from coarse detection to fine-grained structure analysis. Typically, the object structure can be represented as a set of pre-defined keypoints, *e.g.*, shoulder, elbow, or tail and root. As an important step for object structure analysis, general object keypoint localization has attracted increasing attention. Existing keypoint localization methods typically utilize powerful neural networks, *e.g.*, Convolutional Neural Network (CNN) [28, 35] or Vision Transformer (ViT) [21, 38] to either directly regress keypoint coordinates [15] or estimate the keypoint heatmaps [28, 35] to perform localization. Those methods learn keypoint priors purely from training set, *i.e.*, training backbone from scratch on a specific keypoint localization dataset and test it on corresponding test set. This pipeline is limited to the narrow domain defined by the training data, making those methods hard to generalize to different domains or various objects, resulting in poor generalization capability.

Designing a generalizable keypoint localization method is important but challenging. The generalization capability of neural network is inseparable from large scale training data, which is demonstrated by recent generative models in language and vision area, *e.g.*, Large Language Model (LLM) [2] and Latent Diffusion Model (LDM) [25] show remarkable generation performance across a wide range of domains. Therefore a straightforward way to pursue generalization capability is to collect large scale data on a specific task and train model. For instance, SAM [12] and Depth Anything [39] involve a training-annotation-loop to automatically generate image-mask/depth pair to train a powerful segmentation/depth estimator. These methods are effective but not efficient and will consume large resources.

Different from those efforts, this work is motivated to exploit whether a generative model trained on large scale data can be transferred to generalizable keypoint localization task. In depth estimation task, Marigold [9] proposes to fine-tune the U-Net of LDM to generate depth map from noise and input image. This pipeline is efficient and effective as it only requires a small amount of labeled data to achieve satisfied performance. However, directly transferring this pipeline into object keypoint localization is not trivial. Different from depth map containing only single channel, keypoint heatmaps contains more channels than image to reflect keypoint type information, making it infeasible to fine-tune U-Net to get desired outputs with a frozen VAE decoder. Therefore how to transfer LDM into keypoint localization is still a challenging problem.

To tackle this challenge, as well as to achieve generalizable keypoint localization, we utilizes the controllability of LDM to get desired keypoint heatmaps, instead of directly fine-tuning LDM U-Net to output target. Specifically, current LDM performs image generation under the condition guidance such as text, image or semantic maps, the condition plays an important role in generation model to inject the semantic information. As shown in Fig. 1, pre-trained LDM can produce different response map for different input condition. This motivates us to encode the keypoint type information into condition and extract the response map through the controllable mechanism of LDM to exploit the generative priors for keypoint localization.

The above intuition leads to our proposed GenLoc, a generalizable object keypoint localization method. GenLoc performs keypoint localization by exploiting the correlation map between keypoint condition and image latent feature, which is extracted by the Condition-guided Keypoint Heatmap Extraction (CKHE) module. To fully utilize the property of keypoint distribution in the extracted correlation map, we also introduce the Heatmap Expectation Loss (HEL) to effectively optimize the whole model in an end-to-end manner. As shown in Fig. 1, compared with related work [3, 10, 24, 44] on optimizing cross attention map in LDM to find targets, GenLoc models the keypoint distribution over image spatial dimension, which can produce more localized heatmap and achieve better performance.

Our proposed GenLoc can effectively leverage the potential of generative prior in pre-trained image generation model, thus achieving superior performance in various object keypoint localization benchmarks. GenLoc achieves 80.1% AP on animal keypoint localization, outperforming previous SoTA model by 2.3%. Moreover, exploiting generative priors makes GenLoc a data-efficient method. For instance, in few-shot setting, GenLoc achieves 62.8% AP with only 5 images per category, which is higher than previous ScarceNet [14] by 9.5%. In zero-shot keypoint localization setting, GenLoc also shows superior performance.

In summary, GenLoc performs keypoint localization by extracting the correlation map between image feature and keypoint condition. This design can effectively exploit the generative priors in pre-trained LDM and make GenLoc a data-efficient generalizable keypoint localization method. Benefited by the proposed CKHE and HEL, GenLoc achieves superior performance on various keypoint localization datasets, especially in cross-domain, few-shot and zero-shot settings.

2. Related Work

2.1. Keypoint Localization

keypoint localization aims to locate the keypoints of various objects from input RGB images, and existing keypoint localization methods can be divided into regression-based and heatmap-based methods.

Regression-based keypoint localization directly outputs keypoint coordinates via powerful neural network [1, 32]. Due to the high non-linear mapping between input image and numerical coordinates, many works have been proposed to improve the performance of direct regression, including: i) regressing via integration, such as Soft-armgax [29] and Sampling-argmax [16], ii) proposing proposing new loss functions like RLE [15], and iii) designing more powerful backbones to improve the performance of direct regression, such as TokenPose [17] and PETR [26].

Heatmap-based keypoint localization encodes keypoint location with a probability map [31]. This type of methods estimates heatmaps and retrieves keypoint coordinates with a post-processing operation. Currently, heatmap-based methods dominate the field of keypoint localization because heatmap is easy to learn by CNN or Vision Transformer. Pioneer works [22, 28, 35, 38] design powerful CNN or Transformer models to estimate high resolution heatmaps for human/animal/face keypoint localization.

All above methods train model on a narrow domain training data, which leads to limited generalization capability. In contrast, our method exploits generative priors encoded in LDM from large scale data, making it more generalizable to detect keypoints from various image domain and objects.

2.2. Generative Models for Visual Perception

Generative model, *e.g.*, Latent Diffusion Model (LDM) [25] aims to reconstruct data, *e.g.*, image, from random noise without manual annotation, making them effectively utilize large scale unlabeled data to enhance its generalization capability. Therefore, researchers try to transfer the generative models into other areas such as visual perception to boost performance. A line of work utilizes the architecture of generative model, *e.g.*, diffusion model to effectively capture the task-specific data distribution. DDP [6] adopts the multi-step denoising framework into segmentation and depth estimation task and achieves superior performance over conventional architecture. SODA [5] adopts the diffusion framework for representation learning. However, directly transferring generative architecture cannot fully utilize the generative priors trained from large scale data, these methods still have limited generalization capability.

Another line of work exploits how to transfer the generative model with pre-trained weights into downstream tasks. Directly fine-tuning LDM on specific task is a way to achieve this goal. Marigold [9] fine-tunes the U-Net of LDM to estimate the depth map, achieving superior generalization performance in depth estimation. DMP [13] adopts LoRA [4] to fine-tune the U-Net for dense prediction tasks. GED [45] simplifies the multiple denoising steps in LDM by directly predicting the latent edge map. Recently some methods show that optimizing the condition of LDM can also get desired outputs. SLiMe [10] optimizes the condition to perform few-shot segmentation. Diff-Tracker [44] learns prompts to represent object for tracking. Hedlin *et al.* [3] optimizes text embedding for unsupervised learning of keypoints and landmarks. However, these methods simply adopt cross attention map to find targets, which is not effective in keypoint localization. As demonstrated in Sec. 4, optimizing cross attention map like Hedlin *et al.* [3] cannot achieve satisfied performance in standard keypoint localization tasks. Our proposed method can accurately model the keypoint distribution, thus achieves better performance.

3. Method

3.1. Overview

The object structure can be represented by a set of keypoints $\{\mathcal{K}_i\}_{i=1}^n$, where $\mathcal{K}_i$ denotes the coordinates of the i-th type keypoint, *e.g.*, person shoulder or animal tail root, n is the total type of keypoint defined in this object. The goal of object keypoint localization is to estimate the coordinates of target keypoints from input image, *i.e.*,

$$\{\mathcal{K}_i\}_{i=1}^n = \mathrm{locate}(\mathcal{I}), \tag{1}$$

where $\mathcal{I}$ denotes the input image contains a specific object.

We adopt the heatmap technology [31] to locate keypoint, which encodes keypoint into a 2D probability distribution map $\mathcal{H}_i$ and estimates it with a neural network, *i.e.*,

$$\{\mathcal{H}_i\}_{i=1}^n = \Phi(\mathcal{I}), \tag{2}$$

where $\Phi(\cdot)$ can be implemented by Convolutional Neural Network (CNN) [28, 35] or Vision Transformer [38]. Once the model is learned, keypoints can be decoded by finding the maxima in each channel of heatmap $\mathcal{H}$, *i.e.*, $\mathcal{K}_i = \arg\max(\mathcal{H}_i)$. Typical object keypoint localization methods learn the model $\Phi(\cdot)$ purely from training data, *i.e.*, training on a specific domain. This pipeline makes them restricted to a specific image domain or object category and leads to inferior performance and generalization capability.

Instead of learning $\Phi(\cdot)$ from a specific domain, we aims to exploits the rich prior in image generative models to perform localization. Image generative models, *e.g.*, latent diffusion model (LDM) [25], trained on large scale image data have shown promising performance in understanding scene structure and are successfully transferred to depth estimation [9] or edge detection [45] tasks. This motivates us to learn a generalizable localization model $\Phi(\cdot)$ by exploiting the generative priors from LDM. Given an input image $\mathcal{I}$, LDM adopts a pre-trained VAE encoder $\mathcal{E}$ to encoder $\mathcal{I}$ into a latent vector $\mathbf{z} = \mathcal{E}(\mathcal{I})$, from which the original data can be reconstructed via a decoder $\mathcal{D}$ by $\bar{\mathcal{I}} = \mathcal{D}(\mathbf{z}) \approx \mathcal{I}$. LDM performs generation in latent space, *i.e.*, training a denoising U-Net $\mathcal{U}$ to predict the noise from current input z_t, time step t and condition $\mathbf{c}$, *i.e.*,

$$\epsilon = \mathcal{U}(\mathbf{z}_t, t, \mathbf{c}). \tag{3}$$

LDM freezes $\mathcal{E}(\cdot)$ and only trains $\mathcal{U}(\cdot)$ by minimizing the ℓ_2 distance between the predicted noise with its ground truth. Therefore, the generative priors are embedded in $\mathcal{U}(\cdot)$. However, how to extract such information from LDM remains a problem. Previous work [9, 45] shows that directly fine-tune $\mathcal{U}(\cdot)$ on specific tasks can achieve satisfied performance, but directly transferring this approach to keypoint localization is not trivial. Keypoint localization needs to output multiple types of keypoints for single object, which is larger than image channel, making it infeasible to adopt decoder $\mathcal{D}(\cdot)$ to recover latent feature $\mathbf{z}$ to image-level representation.

We propose the GenLoc to extract the generative prior in $\mathcal{U}(\cdot)$ to implement $\Phi(\cdot)$ and perform generalizable keypoint localization. Instead of fine-tuning $\mathcal{U}(\cdot)$ to output corresponding heatmaps, GenLoc exploits the correlation map $\mathcal{C}_i$ between latent feature map $\mathbf{z}$ with condition $\mathbf{c}_i$ to generate final heatmap, this is achieved by the proposed Condition-guided Keypoint Heatmap Extraction (CKHE) module, *i.e.*,

$$\{\mathcal{C}_i\}_{i=1}^n = \mathrm{CKHE}(\mathcal{U}(\cdot), \mathbf{z}, \{\mathbf{c}_i\}_{i=1}^n). \tag{4}$$

In this way, keypoint type is encoded into condition $\mathbf{c}_i$, allowing us flexibly to locate arbitrary type of keypoints.

We train CKHE module with our proposed Heatmap Expectation Loss (HEL), which computes the loss by taking the extracted correlation map $\{\mathcal{C}_i\}_{i=1}^n$ and its corresponding ground truth heatmap $\{\mathcal{H}_i^*\}_{i=1}^n$ as inputs. The computed

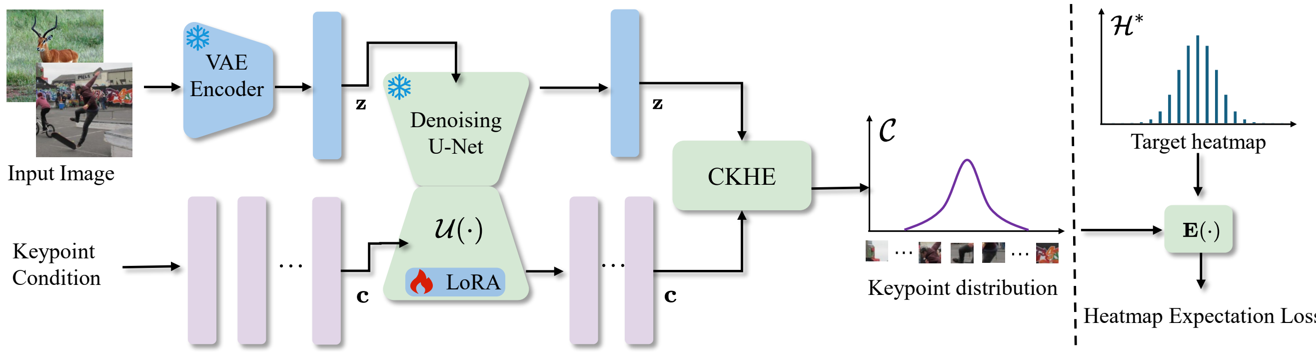

Figure 2. The proposed GenLoc for generalizable keypoint localization by exploiting generative priors from LDM [25]. GenLoc takes object image as input and initializes a set of keypoint condition to extract generative priors. The image input is processed by a frozen VAE encoder to get the latent image feature. The denoising U-Net takes the latent feature and keypoint condition as input and generates corresponding updated output features. Latent feature and keypoint condition are passed to our proposed CKHE module to obtain the keypoint distribution map. The whole pipeline is optimized by our proposed Heatmap Expectation Loss in an end-to-end manner.

loss $\mathcal{L}_{HEL}$ can be represented as,

$$\mathcal{L}_{\text{HEL}} = \sum_{i=1}^{n} \mathcal{D}(\mathcal{C}_i, \mathcal{H}_i^*), \tag{5}$$

where $\mathcal{D}(\cdot)$ computes the loss by comparing the difference between the extracted correlation map and its ground truth.

We supervise CKHE with HEL to perform localization, Fig. 2 shows the pipeline of the proposed GenLoc. During training we freeze the VAE encoder $\mathcal{E}$, U-Net $\mathcal{U}(\cdot)$ and only adjust condition $\mathbf{c}$ and a small set of LoRA [4] parameters to extract the generative prior. This strategy allows us to keep the rich knowledge in $\mathcal{U}(\cdot)$ and accurately locate the target keypoints. Compared with directly fine-tuning $\mathcal{U}(\cdot)$ in previous work, our method GenLoc is more effective in maintaining and exploiting the generative prior for keypoint localization. Implementations to CKHE and HEL can be found in the following parts.

3.2. Localization from Generative Priors

Given an input image $\mathcal{I}$ and pre-trained LDM U-Net $\mathcal{U}(\cdot)$, GenLoc aims to optimize a set of keypoint condition $\mathbf{c} = \{\mathbf{c}_i\}_{i=1}^{n}$ to locate the target keypoint coordinates. The core components of GenLoc are the Condition-guided Keypoint Heatmap Extraction (CKHE) module to extract generative prior from $\mathcal{U}(\cdot)$ and Heatmap Expectation Loss (HEL) to supervise the whole model training.

Condition-guided Keypoint Heatmap Extraction. Different from previous work that direct fine-tuning U-Net to estimate heatmap, we exploit the controllable generation mechanism in LDM for heatmap generation. Current most generative models implement controllable mechanism via condition input, *e.g.*, text or image. Condition is injected into generation process via the widely adopted attention mechanism in Transformer [33]. Specifically, given the latent feature $\mathbf{z} \in \mathbb{R}^{m \times d}$ and condition embedding $\mathbf{c} \in \mathbb{R}^{n \times d}$, we pass them to $\mathcal{U}(\cdot)$ to get the updated latent feature and condition, the cross attention can be computed by,

$$\mathcal{S} = \underset{\mathbf{n}}{\text{softmax}}(\frac{\mathbf{z} \cdot \mathbf{c}^T}{\sqrt{d}}), \tag{6}$$

where $\mathcal{S} \in \mathbb{R}^{m \times n}$ denotes the attention score between latent feature $\mathbf{z}$ and condition $\mathbf{c}$, m and n denotes the number of image tokens and keypoint condition embeddings, d indicates the dimension of features. Note that the softmax is performed along different keypoint conditions, this indicates that each token from latent $\mathbf{z}$ independently finds the similar keypoint condition and aggregates its feature. An illustration of cross attention in LDM is shown in Fig. 3.

Although $\mathcal{S}$ in Eq. (6) can be interpreted as heatmap to perform localization, directly adopting it has several drawbacks. First, the similarity computation in $\mathcal{S}$ is independent among m image latent tokens, which does not satisfy the property of heatmap that only has high response at target spatial position, leading to inaccurate keypoint distribution estimation. Second, different conditions are correlated in $\mathcal{S}$ and it is hard to differentiate spatial adjacent keypoints as each image token can only produce high response to only one type of keypoint. Therefore we propose the CKHE to avoid above limitation. Different from cross attention in Eq. (6), CKHE considers to compute correlation map between latent feature $\mathbf{z}$ and keypoint condition $\mathbf{c}$ such that each token in $\mathbf{z}$ is correlated but different condition in $\mathbf{c}$ is independent. To achieve this goal, we exchange the computation order in Eq. (6) and compute the softmax over latent tokens, not keypoint condition, *i.e.*,

$$\mathcal{C} = \underset{\mathbf{m}}{\text{softmax}}(\frac{\mathbf{c} \cdot \mathbf{z}^T}{\sqrt{d}}), \tag{7}$$

where $\mathcal{C} \in \mathbb{R}^{n \times m}$ denotes the estimated heatmap by CKHE. Besides Eq. (7), we also need to fine-tune corresponding attention layer parameters to adapt the changed computation

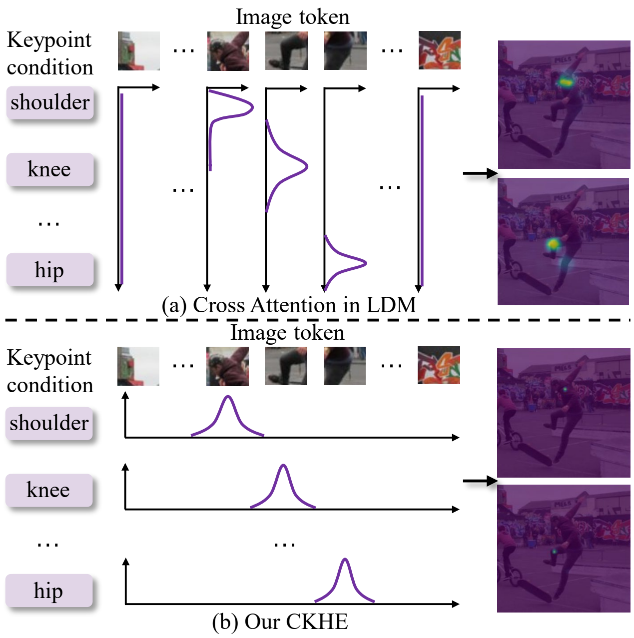

Figure 3. Illustration the different interaction mechanism between image latent feature and keypoint condition. (a) LDM adopts cross attention to aggregate condition information into latent feature, which is not suitable for keypoint localization and leads to inferior heatmap results. (b) Our proposed CKHE models the keypoint distribution over image latent features and decouple the keypoint condition, which generates more localized heatmaps and can differentiate spatial adjacent keypoints.

way, which is implemented by LoRA. Note that although Eq. (7) is similar to Eq. (6), they are different in many aspects. As state above, Eq. (7) adopts keypoint condition as query to compute similarity score, therefore the tokens in latent $\mathbf{z}$ are naturally correlated, namely only a small number of tokens have high probability in the estimated correlation map $\mathcal{C}$. Moreover, Eq. (7) also de-correlates the keypoint condition, making it feasible to detect spatial adjacent keypoints. As shown in Fig. 3, CKHE can model keypoint distribution more accurately and generate more localized heatmap, achieving better performance.

Heatmap Expectation Loss. We aims to optimize the condition $\mathbf{c}$ so that the obtained correlation map $\mathcal{C}$ can be used to accurately locate the target keypoint. To achieve this goal we compute loss by measuring the difference between $\mathcal{C}$ and target ground truth $\mathcal{H}^*$. A straightforward approach is to adopt the widely used ℓ_2 distance, *i.e.*,

$$\mathcal{L}_{\text{MSE}} = \ell_2(\mathcal{C}_i, \mathcal{H}_i^*). \tag{8}$$

However, adding element-wise supervision in Eq. (8) is not effective because $\mathcal{H}^*$ is manually defined unnormalized Gaussian distribution with fixed variance, which cannot reflect the underlying true distribution and will hinder the model to learn a accurate distribution to model the target keypoint, which is also observed in previous work [15, 20].

Although the underlying distribution of target keypoint is unknown, it should have large overlap with $\mathcal{H}^*$ to produce high response at the target location. Therefore we instead model the overlap between $\mathcal{C}$ and $\mathcal{H}^*$ and propose a Heatmap Expectation Loss (HEL) to address above limitation. Inspired by Integral Pose [29], we compute the expectation score by integral the target $\mathcal{H}_i^*$ over the estimated distribution $\mathcal{C}_i$ to model the overlap between $\mathcal{C}$ and $\mathcal{H}^*$. The goal of HEL is to maximize the expectation score, *i.e.*,

$$\mathcal{L}_{\text{HEL}} = -\log\left(\sum_{j=1}^{m} \mathcal{C}_i[j] \cdot \mathcal{H}_i^*[j]\right), \tag{9}$$

where m denotes the number of image tokens. As illustrated in Fig. 2 right, higher expectation score indicates larger overlap between $\mathcal{C}_i$ and $\mathcal{H}_i^*$, leading to sharper heatmap in Fig. 3. Compared with $\mathcal{L}_{\text{MSE}}$, $\mathcal{L}_{\text{HEL}}$ can fully utilize the normalization property of $\mathcal{C}$ to producing more effective gradients during training, thus achieves better performance.

To better show the advantage of the proposed HEL, we give a detailed gradient analysis of Eq. 8 and Eq. 9 and show how them different in optimizing the correlation map $\mathcal{C}$ to its target $\mathcal{H}^*$. Specifically, the gradient of the loss function to each image latent token $\mathcal{C}_i[j] \in \mathcal{C}_i$ can be computed as,

$$\frac{\partial \mathcal{L}(\mathcal{C}_i[j], \mathcal{H}_i^*[j])}{\partial \mathcal{C}_i[j]} = \begin{cases} -2(\mathcal{H}_i^*[j] - \mathcal{C}_i[j]) & \text{for } \mathcal{L}_{\text{MSE}} \\ -\mathcal{H}_i^*[j]/(\sum_j \mathcal{C}_i[j] \cdot \mathcal{H}_i^*[j]) & \text{for } \mathcal{L}_{\text{HEL}} \end{cases} \tag{10}$$

For positive samples in $\mathcal{C}_i$ (*i.e.* $\mathcal{H}_i^*[j] = 1$), the gradient of HEL and MSE loss is small when $\mathcal{C}_i[j] \to 1$. Their difference is reflected on hard samples. When a positive position is predicted with a low probability, HEL will produce larger gradient than MSE loss for the model training, thus is more effective for model training. This strategy is also adopted in other loss functions such as focal loss in object detection and triplet loss with hard mining in metric learning.

3.3. Discussions

Compared with conventional keypoint localization methods trained on a specific domain, GenLoc exploits the rich generative priors in LDM, which is pre-trained on large scale image dataset. Equipped with our proposed CKHE and HEL, GenLoc can accurately locate keypoint and show superior performance on various localization benchmarks.

GenLoc share some similarity with several works like SLiMe [10], Diff-Tracker [44] and Hedlin *et al.* [3] in optimizing condition to find desired targets. However, as illustrated in CKHE, all previous methods directly optimize cross attention map to find targets, which is not suitable for keypoint localization, as it requires accurately modeling the keypoint distribution over spatial dimension. Our proposed CKHE and HEL boost the cross attention baseline by a large margin, producing more localized heatmap in Fig 3. More quantitative results can be found in Sec. 4.

4. Experiments

4.1. Datasets and Evaluation Metrics

We evaluate GenLoc on several widely used keypoint localization benchmarks, *e.g.*, COCO Keypoint [19] and Human-Art [8] for human keypoint localization, AP-10K [40] for animal keypoint localization.

COCO Keypoint [19] is a large scale human keypoint localization dataset with 17 annotated keypoints, *e.g.*, nose, left ear, *etc*. It contains more than 200,000 images and 250,000 person instances. COCO Keypoint is divided into three parts: 57k images for training set, 5k images for validation set and 20k images for test-dev set.

Human-Art [8] is a large scale human keypoint localization containing both natural scenes such as sports or outdoor, and artificial scenes including cartoon, digital art, ink painting and *etc*. We adopt it to evaluate the cross-domain localization performance of different methods. Similar to COCO, Human-Art is also annotated with 17 keypoints and contains 35k images for training set and 5k images for validation set. In this paper we only adopt its validation set.

AP-10K [40] is a large scale animal keypoint localization dataset contains 10,015 images collected and filtered from 23 animal families and 54 species. 17 keypoints are annotated in the dataset. AP-10K splits the train, validation, and test sets with a ratio of 7:1:2 per animal species.

Evaluation metrics. We follow the standard evaluation metric and use OKS-based metrics for keypoint localization. We report average precision with different thresholds and object sizes: AP, AP^{50}, AP^{75}, AP^{M} and AP^{L}.

4.2. Implementation Details

All experiments are implemented based on PyTorch [23]. We utilize Stable Diffusion v1.5 [25] as the pre-trained image generative model. The dimension of condition embedding of SD1.5 is set to 768. We additionally fine-tune LDM by LoRA, with hidden dimension $r = 16$. We insert CKHE into all cross attention layers in U-Net up-blocks. The condition embedding and LoRA modules are optimized by Adam optimier [11] with 1e-2 and 1e-3 learning rate, respectively. We train GenLoc with batch size of 200 through gradient accumulation and 10 epochs on AP-10k for ablation study. For final comparison, we train model 100 epochs on AP-10K and 14 epochs for COCO Keypoints.

4.3. Ablation Study

This section aims to investigate the contribution of each proposed components in GenLoc, including the CKHE module and proposed Heatmap Expectation Loss. We also present training analysis in our method.

Effectiveness of CKHE. We first show the effectiveness of the proposed CKHE on keypoint localization. The results are shown in the top rows of Table 1. LDM adopts

Component	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Heatmap generation					
Cross Attention	60.9	92.3	65.8	49.7	61.2
CKHE	**70.7**	**94.3**	**78.5**	**53.4**	**71.0**
Loss function					
MSE Loss	70.7	94.3	78.5	53.4	71.0
HEL	**73.3**	**96.8**	**80.9**	**55.1**	**73.7**
Fine-tuning strategy					
Only condition	60.1	93.1	64.6	37.9	60.7
+ LoRA	**73.3**	**96.8**	**80.9**	**55.1**	**73.7**
Noise time step selection					
High noise($t = 601$)	15.4	41.3	8.5	7.0	15.7
Low noise($t = 21$)	71.3	94.9	79.0	57.1	71.6
No noise	**73.3**	**96.8**	**80.9**	**55.1**	**73.7**

Table 1. Ablation study on each proposed component on AP-10K.

cross attention to aggregate the information from condition to image latent feature, which is inconsistent with the goal of keypoint localization. In contrast, our proposed CKHE regards condition as query and computes the condition and image latent feature along spatial token dimension, thus can better estimate the distribution of keypoints. As shown in Table 1, CKHE significantly outperforms cross attention by a large margin under the same setting, *e.g.*, higher than 9.8%. In Fig. 3 we also can observe that CKHE produces more localized heatmaps, leading to better performance.

Effectiveness of HEL. In the middle rows of Table 1 we also show the effectiveness of the HEL over traditional MSE loss. As stated in Sec. 3, MSE loss is not effective in learning accurate keypoint distribution from the manually defined groud truth heatmap. Our proposed HEL can fully utilize the normalization property of correlation map and produce more effective gradient to train model. As shown in Table 1, HEL outperforms MSE loss by 2.6%, indicating its advantages in learning accurate keypoint distribution.

Fine-tuning with LoRA. We also verify the necessity of fine-tuning the model with LoRA. Previous methods [3, 10, 44] only optimizing condition to find targets, which cannot achieve superior performance in keypoint localization, *i.e.*, only obtaining 60.1%. This may be that LDM does not need a well-localized correlation map for image generation, which is exactly required by keypoint localization. By fine-tuning LDM with LoRA, we achieve a balance between the localization performance and the generalization capability, outperforming baseline by 13.2% AP.

Noise time step selection. LDM performs generation from noisy input and previous work [30] extracts features from U-Net by adding noise to input image. Therefore we investigate how to add noise to input image for keypoint localization. We evaluate model performance under high

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
SimplePose [35]	69.9	-	-	-	-
Hourglass [22]	72.9	-	-	-	-
HRNet-W32 [28]	72.2	93.5	78.9	55.7	72.9
HRNet-W48 [28]	72.8	93.6	80.2	57.7	73.5
ViPNAS [37]	67.1	-	-	-	-
HRFormer-S [41]	71.7	-	-	-	-
HRFormer-B [41]	73.5	-	-	-	-
RTMPose [7]	72.2	93.9	78.8	56.9	72.8
CLAMP-B [43]	74.3	-	-	-	-
CLAMP-L [43]	77.8	-	-	-	-
Ours	**80.1**	**97.3**	**87.6**	**59.3**	**80.5**

Table 2. Comparison on AP-10K in animal keypoint localization.

noisy, low noisy and clean input (namely no noise is added to image), and results are shown in the bottom of Table 1. It can be observed that a low noise is preferred by model to extract image feature and we find that adopting clean image as input achieves the best performance. Therefore in the following experiments we adopt clean image as input.

4.4. General Object Keypoint Localization

The section aims to demonstrate the effectiveness of the proposed GenLoc on various object keypoint localization tasks. We compare GenLoc with recent methods on animal keypoint localization benchmark AP-10K and human keypoint localization benchmark COCO Keypoint. The results are shown in Table 2 and Table 3.

Animal Keypoint Localization. We first compare GenLoc with other methods on animal keypoint localization task and report performance on AP-10K. As shown in Table 2, existing animal keypoint localization methods are belong to heatmap-based method, *e.g.*, SimplePose [35], HRNet [28] and HRFormer [41]. Compared with these conventional methods, GenLoc achieves superior performance, *e.g.*, 80.1% AP on AP-10K, which is higher than HRFormer by a large margin. Recent CLAMP [43] embeds language information into animal keypoint localization, which improves the localization performance. Compared with CLAMP, our method can fully utilize the large LDM pre-trained on large scale image data, outperforming CLAMP by 2.3% AP. All this indicates the superior performance of GenLoc on animal keypoint localization task.

Human Keypoint Localization. We then show that GenLoc can also locate person keypoint and compare on COCO Keypoint. As shown in Table 3, existing methods can be divided into heatmap-based method: SimplePose [35], HRNet [28] and ViTPose [38], and regression-based methods: DeepPose [32] and RLE [15]. All of them learn the keypoint prior purely from training data. Compared with these works, our method can exploit the generative priors in LDM and achieving better performance.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
DeepPose [32]	53.8	82.6	59.2	52.2	57.3	66.8
RLE [15]	74.0	91.5	81.6	70.9	78.5	76.8
SimplePose [35]	74.4	92.6	82.5	71.5	79.2	77.6
HRNet [28]	76.8	93.6	83.6	74.0	81.5	79.6
SimCC [18]	76.5	93.2	83.1	73.6	81.5	79.7
ViTPose [38]	77.4	93.6	84.8	74.7	81.9	80.2
Ours	**77.6**	**94.6**	**85.7**	**74.8**	**82.6**	**80.7**

Table 3. Comparison on COCO in human keypoint localization.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
SimplePose [35]	48.4	73.0	50.7	27.2	50.7	52.8
SimCC [18]	51.7	75.2	54.8	26.2	54.3	57.0
HRNet [28]	53.4	76.3	56.5	30.4	55.9	57.5
ViTPose [38]	53.8	77.9	57.4	31.4	56.6	58.7
LocLLM [34]	64.8	87.4	70.4	40.9	67.4	69.3
Ours	**67.0**	**87.9**	**72.5**	**43.2**	**69.9**	**70.8**

Table 4. Comparison on Human-Art in cross-domain keypoint localization. All models are trained on COCO Keypoint.

Method	5	10	15	20	25
Labeled data + additional unlabeled data					
UDA [36]	0.429	0.519	0.566	0.580	0.628
FixMatch [27]	0.478	0.544	0.589	0.601	0.631
FlexMatch [42]	0.466	0.555	0.596	0.618	0.646
ScarceNet [14]	0.533	0.597	0.632	0.654	0.681
Only labeled data					
HRNet [28]	0.360	0.463	0.511	0.547	0.588
Ours	**0.628**	**0.668**	**0.688**	**0.692**	**0.711**

Table 5. Comparison on AP-10K in few-shot localization. We train model when 5, 10, 15, 20, 25 images per species are labeled.

4.5. Cross-Domain Keypoint Localization

In this section we aims to demonstrate the generalization capability of GenLoc by comparing with other methods in cross-domain keypoint localization, *i.e.*, training model on one dataset and test it on other domain dataset without finetuning. Human-Art [8] dataset contains different image domain from COCO Keypoint, therefore we adopt it as test set to evaluate models trained on COCO Keypoint.

As shown in Table 4, although previous methods achieve superior performance in COCO Keypoint, they obtain inferior performance over Human-Art benchmarks. This is because limited training data in COCO Keypoint cannot provide rich prior for these model to locate keypoint from different image domain. Our method utilizes the prior in pretrained LDM, which is trained on large scale data to ensure its domain diversity. It achieves 67.0% AP on Human-Art, outperforming previous methods.

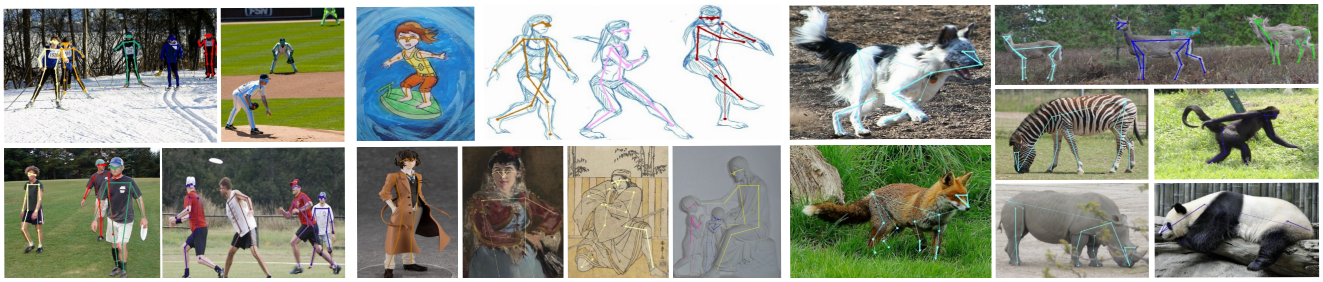

Figure 4. Visualization of keypoint localization results from COCO Keypoint (left), Human-Art (middle) and AP-10K (right).

Method	Train	Test	AP	AP_{50}	AP_{75}	AR
SimplePose [35]	Bovidae	Canidae	41.3	79.4	36.4	49.1
CLAMP [43]	Bovidae	Canidae	46.9	84.4	45.6	53.8
Ours	Bovidae	Canidae	**72.8**	**97.4**	**82.4**	**75.8**
SimplePose [35]	Bovidae	Equidae	41.9	71.8	40.3	46.6
CLAMP [43]	Bovidae	Equidae	46.6	75.6	47.5	51.2
Ours	Bovidae	Equidae	**76.0**	**95.7**	**83.1**	**80.7**
SimplePose [35]	Bovidae	Felidae	22.0	52.4	15.0	28.4
CLAMP [43]	Bovidae	Felidae	28.7	67.6	18.9	36.3
Ours	Bovidae	Felidae	**64.3**	**94.7**	**69.2**	**69.2**
SimplePose [35]	Canidae	Felidae	39.6	74.1	34.5	46.6
CLAMP [43]	Canidae	Felidae	48.4	85.7	44.0	55.1
Ours	Canidae	Felidae	**73.4**	**95.8**	**80.9**	**77.4**
SimplePose [35]	Canidae	Cricetidae	16.1	41.7	10.1	26.0
CLAMP [43]	Canidae	Cricetidae	22.0	51.1	14.1	31.6
Ours	Canidae	Cricetidae	**50.2**	**84.2**	**51.0**	**55.8**
SimplePose [35]	Canidae	Equidae	20.5	43.9	16.6	25.3
CLAMP [43]	Canidae	Equidae	28.4	59.1	23.1	34.1
Ours	Canidae	Equidae	**61.2**	**85.4**	**61.5**	**66.0**

Table 6. Comparison on AP-10K in zero-shot localization.

4.6. Few-shot Keypoint Localization

In this section we aims to show that GenLoc is a highly data-efficient generalizable keypoint localization method, *e.g.*, it only requires a small set of labeled data to achieve satisfied performance. We follow the ScarceNet [14] setting to randomly select $5, 10, 15, 20, 25$ labeled images of each species from original AP-10K training set to train the model, and directly test it on the complete test set.

As shown in Table 5, existing methods in few-shot setting can be divided into two categories. The first is semi-supervised methods, *e.g.*, UDA [36] and ScarceNet [14], these methods require additional unlabeled data to boost the model performance. The second kind of methods only use labeled data to train model like HRNet [28]. However, limited training data cannot provide rich prior to perform keypoint localization, thus HRNet achieves much worse performance than semi-supervised methods like ScarceNet. Different from previous methods, our method exploits the rich generative priors from image generation model to perform keypoint localization, and achieves superior performance over semi-supervised methods when only using labeled data. For instance, GenLoc achieves 62.8% AP in 5-shot setting, which is higher than ScarceNet by 9.5% AP and HRNet by 26.8% AP, respectively.

4.7. Zero-shot Keypoint Localization

In the final section we further evaluate the generalization capability of GenLoc on unseen objects in zero-shot keypoint localization. We follow CLAMP [43] setting to conduct experiments by testing model on unseen animal species on AP-10K, *e.g.*, training model on one animal species and test it on another species. This setting is more challenge because the appearance, size and background environment are quite different for different animal species.

We follow CLAMP [43] to conduct different zero-shot keypoint localization experiments according to whether the animal during training and test belong to the same animal order or not. As shown in Table 6, we compare our method with previous methods SimplePose [35] and CLAMP [43]. It can be observed that our method significantly outperform these methods by a large margin. For instance, when training on Bovidae specie and testing on Canidae specie, our GenLoc achieves 72.8% AP, which is higher than CLAMP by 25.9% AP. When training and testing species belong to different orders, *e.g.*, training on Canidae and testing on Felidae, our method can get more improvement, outperforming CLAMP by 35.6% AP. Such large improvement comes from the rich generative priors from large scale data pretrained LDM, which can provide more prior than a narrow domain training data. Our method can fully utilize the generative prior to achieve generalizable keypoint localization in zero-shot setting.

5. Conclusion

This paper proposes GenLoc, a generalizable object keypoint localization method. GenLoc exploits the generative priors in image generation model to perform keypoint localization. It first adopts the proposed CKHE to extract the correlation map between condition and latent image feature, then optimizes the correlation map by our Heatmap Expectation Loss to perform localization. Experiments on various keypoint localization benchmarks demonstrate the strong generalization capability of the proposed method.

Acknowledgement

This work is supported in part by Guanghua Talent Project of Southwestern University of Finance and Economics and Sichuan Science Foundation project (Grants No. 2024ZDZX0002 and 2024NSFTD0054), in part by CCF-Baidu Open Fund (No. 202422) and open research fund of Key Laboratory of Machine Intelligence and System Control, Ministry of Education (No. MISC-202406), in part by Grant No. 2023-JCJQ-LA-001-088, in part by Natural Science Foundation of China under Grant No. U20B2052, 61936011, 62402013, in part by the Okawa Foundation Research Award, in part by the Ant Group Research Fund, and in part by the Kunpeng&Ascend Center of Excellence, Peking University.

References

[1] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *CVPR*, 2016. 2

[2] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 1

[3] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Xingzhe He, Hossam Isack, Abhishek Kar, Helge Rhodin, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised keypoints from pretrained diffusion models. In *CVPR*, 2024. 2, 3, 5, 6

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3, 4

[5] Drew A Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K Lampinen, Andrew Jaegle, James L McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. Soda: Bottleneck diffusion models for representation learning. In *CVPR*, 2024. 3

[6] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. Ddp: Diffusion model for dense visual prediction. In *ICCV*, 2023. 3

[7] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023. 7

[8] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *CVPR*, 2023. 6, 7

[9] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024. 2, 3

[10] Aliasghar Khani, Saeid Asgari Taghanaki, Aditya Sanghi, Ali Mahdavi Amiri, and Ghassan Hamarneh. Slime: Segment like me. *arXiv preprint arXiv:2309.03179*, 2023. 2, 3, 5, 6

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[13] Hsin-Ying Lee, Hung-Yu Tseng, and Ming-Hsuan Yang. Exploiting diffusion prior for generalizable dense prediction. In *CVPR*, 2024. 3

[14] Chen Li and Gim Hee Lee. Scarcenet: Animal pose estimation with scarce annotations. In *CVPR*, 2023. 2, 7, 8

[15] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *ICCV*, 2021. 1, 2, 5, 7

[16] Jiefeng Li, Tong Chen, Ruiqi Shi, Yujing Lou, Yong-Lu Li, and Cewu Lu. Localization with sampling-argmax. *NeurIPS*, 2021. 2

[17] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *CVPR*, 2021. 2

[18] Yanjie Li, Sen Yang, Peidong Liu, Shoukui Zhang, Yunxiao Wang, Zhicheng Wang, Wankou Yang, and Shu-Tao Xia. Simcc: A simple coordinate classification perspective for human pose estimation. In *ECCV*, 2022. 7

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6

[20] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. In *CVPR*, 2021. 5

[21] Zehong Ma, Shiliang Zhang, Longhui Wei, and Qi Tian. Ovmr: Open-vocabulary recognition with multi-modal references. In *CVPR*, 2024. 1

[22] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 2, 7

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6

[24] Duo Peng, Zhengbo Zhang, Ping Hu, Qiuhong Ke, David KY Yau, and Jun Liu. Harnessing text-to-image diffusion models for category-agnostic pose estimation. In *ECCV*, 2025. 2

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 4, 6

[26] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *CVPR*, 2022. 2

[27] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS*, 2020. 7

[28] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 1, 2, 3, 7, 8

[29] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 2, 5

[30] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *NeurIPS*, 2023. 6

[31] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *NeurIPS*, 2014. 2, 3

[32] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 2, 7

[33] A Vaswani. Attention is all you need. *NeurIPS*, 2017. 4

[34] Dongkai Wang, Shiyu Xuan, and Shiliang Zhang. Locllm: Exploiting generalizable human keypoint localization via large language model. In *CVPR*, 2024. 7

[35] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 1, 2, 3, 7, 8

[36] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *NeurIPS*, 2020. 7, 8

[37] Lumin Xu, Yingda Guan, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Vipnas: Efficient video pose estimation via neural architecture search. In *CVPR*, 2021. 7

[38] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *arXiv preprint arXiv:2204.12484*, 2022. 1, 2, 3, 7

[39] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 2

[40] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. *arXiv preprint arXiv:2108.12617*, 2021. 6

[41] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. *NeurIPS*, 2021. 7

[42] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *NeurIPS*, 2021. 7

[43] Xu Zhang, Wen Wang, Zhe Chen, Yufei Xu, Jing Zhang, and Dacheng Tao. Clamp: Prompt-based contrastive learning for connecting language and animal pose. In *CVPR*, 2023. 1, 7, 8

[44] Zhengbo Zhang, Li Xu, Duo Peng, Hossein Rahmani, and Jun Liu. Diff-tracker: Text-to-image diffusion models are unsupervised trackers. In *ECCV*, 2025. 2, 3, 5, 6

[45] Caixia Zhou, Yaping Huang, Mochu Xiang, Jiahui Ren, Haibin Ling, and Jing Zhang. Generative edge detection with stable diffusion. *arXiv preprint arXiv:2410.03080*, 2024. 3