Harnessing Text-to-Image Diffusion Models for Category-Agnostic Pose Estimation

Duo Peng[1], Zhengbo Zhang[1], Ping Hu[2], Qiuhong Ke[3], David K. Y. Yau[1], and Jun Liu[1,4,⋆]

[1] Singapore University of Technology and Design
[2] University of Electronic Science and Technology of China
[3] Monash University
[4] Lancaster University
{duo_peng, Zhengbo_Zhang}@mymail.sutd.edu.sg chinahuping@gmail.com
Qiuhong.Ke@monash.edu j.liu81@lancaster.ac.uk

Abstract. Category-Agnostic Pose Estimation (CAPE) aims to detect keypoints of an arbitrary unseen category in images, based on several provided examples of that category. This is a challenging task, as the limited data of unseen categories makes it difficult for models to generalize effectively. To address this challenge, previous methods typically train models on a set of predefined base categories with extensive annotations. In this work, we propose to harness rich knowledge in the off-the-shelf text-to-image diffusion model to effectively address CAPE, without training on carefully prepared base categories. To this end, we propose a Prompt Pose Matching (PPM) framework, which learns pseudo prompts corresponding to the keypoints in the provided few-shot examples via the text-to-image diffusion model. These learned pseudo prompts capture semantic information of keypoints, which can then be used to locate the same type of keypoints from images. We also design a Category-shared Prompt Training (CPT) scheme, to further boost our PPM's performance. Extensive experiments demonstrate the efficacy of our approach.

Keywords: Category-Agnostic Pose Estimation · Diffusion Model

1 Introduction

Pose estimation [2, 25, 66] is one of the fundamental tasks in computer vision and has a wide range of real-world applications including AR/VR [3, 13, 23], autonomous driving [10, 17, 63], human care [14, 34, 58], scene understanding [15,32,56], etc. It aims to estimate the locations of keypoints that are pre-defined by humans. In the field of pose estimation, there exist diverse categories such as human body, human face, furniture, vehicles, etc. While existing methods [47, 65] have achieved notably good performance on these categories, they typically focus on training deep neural networks to handle a single category only, which

⋆ Corresponding author

hinders their generalization to real-world scenarios that may contain diverse unseen categories.

To address the generalization problem, recent research [57] has proposed the task of Category-Agnostic Pose Estimation (CAPE). In CAPE, before testing, the model is given one or a few examples (i.e., few-shot examples) of an arbitrary unseen category. After learning on the given few-shot examples, the model is expected to perform inferences on test images of this unseen category, aiming to find pixel locations in test images that share the same semantic meanings and structural characteristics as the keypoints in the provided few-shot example images. In this way, the model can achieve category-agnostic pose estimation with only few-shot examples.

Considering the limited information provided by few-shot examples, existing CAPE methods generally [5, 28, 45, 46, 57] resort to external knowledge to compensate for the scarcity of pose data. They typically train the model on a set of predefined base categories, and then adapt the model to unseen categories based on the given few-shot examples. While being effective, the construction of base categories requires collecting and labeling data of as many categories as possible, which can be labor-intensive and time-consuming. Also, relying solely on one or a few examples to learn for unseen categories, the model trained on base categories still struggles to generalize effectively. Existing efforts in this area [45, 57] have shown that even with large-scale training, these methods may still yield unsatisfactory pose estimation results when tested on unseen categories (e.g., *furniture* and *vehicle*) that differ significantly from the base categories (e.g., *face* and *human body*). These limitations spark the need for a data-efficient and generalization-effective method for CAPE.

Recently, text-to-image diffusion models [31, 36, 43], such as Stable Diffusion [40], have shown impressive performance in generating photo-realistic images in response to user-defined prompts. Given that existing text-to-image diffusion models can generate high-fidelity images that are visually reasonable, we believe that they contain a wealth of knowledge about the object's semantics, structures, compositions, and spatial relations. In other words, for a text-to-image diffusion model to successfully create realistic images of an object, it must know what components make up this object and it also must understand the correct positions of these components. This concept is similarly reflected in [20]. Inspired by this insight, in this work, we propose to harness the power of text-to-image diffusion models to address the CAPE task that involves spatial composition reasoning, by only learning from few-shot examples of unseen categories. In this way, we may not need to carefully prepare labeled data of base categories but can still achieve good results. However, it is a non-trivial problem, as the text-to-image diffusion model focuses on generating images from texts, lacking a mechanism to estimate keypoint locations for images, thus leading to the challenge in utilizing the potential benefits of the diffusion model.

To tackle the above challenge, in this paper, we consider the CAPE task from the perspective of regarding this task as establishing spatial correspondences (mappings) between example images and test images, which is shown in

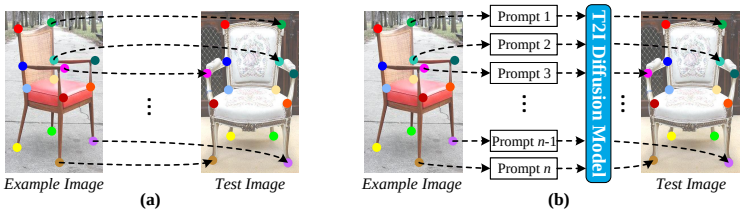

Fig. 1: (a) The CAPE task can be regarded as establishing spatial (keypoint) correspondences between example images and test images. **(b)** To address CAPE, our PPM learns prompts that serve as bridges to build these correspondences.

Fig. 1 (a). This perspective inspires us a feasible way to leverage text-to-image diffusion models to address CAPE, raising a question: *Given that the text-to-image diffusion model essentially builds the correspondences (mappings) between texts and images, can we use such text-image correspondences in diffusion models to establish image-image spatial correspondences for CAPE?* To address this question, we delve deeper into the architecture of text-to-image diffusion models. In particular, we focus on the commonly-used cross-attention mechanism in text-to-image diffusion models, which facilities the visual-textual interactions in the text-to-image generation process. [20] has revealed that the cross-attention maps extracted from text-to-image diffusion models help to highlight the text-related regions of the images. As illustrated in the Fig. 3 (a), given an image of a bicycle and the text prompt "saddle", the cross-attention map highlights the saddle area. Similarly, if the text prompt is changed to "front wheel" or "back wheel", the corresponding regions are highlighted (see Fig. 3 b), showcasing the model's capability to perceive different parts of the object based on the given text input. This is because the diffusion model is trained on massive text-image pairs. After its training, the cross-attention mechanism in the diffusion model is able to understand the structural and spatial information of object compositions in a joint visual-textual manner.

Motivated by this observation, we propose to extend the capabilities of text-to-image diffusion models beyond image generation to address the practical CAPE problem through the following approach: *In CAPE, before testing on images of an unseen category, we are given only one or a few labeled examples of this category. If we can find out what text 'prompt' corresponds to the particular keypoint in the given example image, then the found 'prompt' can be used to drive the text-to-image diffusion model to locate (highlight) a semantically corresponding keypoint in the test image. In this way, the found 'prompt' can serve as a bridge that builds image-to-image correspondences (see Fig. 1 b), thereby harnessing the knowledge in the diffusion model to address the CAPE task.* Following this idea, an intuitive solution is to find an actual text prompt for each keypoint in the given examples. However, the text descriptions of image keypoints are usually not easy to obtain [39], since many keypoints (e.g., keypoints in Figs. 1, 6 and 7) can be difficult to describe in texts even for human beings.

To overcome these challenges, in this paper, we propose a novel framework, which we call Prompt Pose Matching (PPM), to find suitable prompt representations by learning pseudo prompts. Specifically, given few-shot examples of an unseen category, PPM uses the labeled examples to learn (optimize) pseudo prompts, to render them to be semantically aligned with the corresponding keypoints of this unseen category. Learning from the few-shot examples, the pseudo prompts can serve as bridges to find the corresponding keypoints in the test images via cross-attention maps of the diffusion model, thereby establishing the required keypoint correspondences between the example images and the test images. In this way, PPM can work solely based on the few-shot examples of unseen categories, even without training on base categories. Experiments on MP-100 [57] benchmark, which contains over 20K instances covering 100 categories, demonstrate that our PPM solely learning on few-shot examples, performs better than previous methods that require training on extensive data of base categories.

Beyond this core technical contribution, we also design a scheme, which we call Category-shared Prompt Training (CPT) for improved keypoint identification of the PPM. Given that pseudo prompts learned from few-shot examples are category-specific (i.e, customized to the unseen category), they may overlook common pose knowledge shared across different categories, such as the common positional characteristics of keypoints (e.g., keypoints tend to be located at important positions, like corners, edges, or other semantically meaningful regions, regardless of the object's category). Therefore, we further introduce a CPT scheme to acquire a category-shared pseudo prompt to enrich the prompt representation for better keypoint localization. Using CPT, the proposed PPM yields further improved performance, significantly outperforming previous methods.

2 Related Work

Pose Estimation endeavors to localize semantic or interest keypoints, such as human body parts [1,25,50], facial landmarks [2,27,48], hand keypoints [7,8,66], and vehicle poses [38, 44, 64] from the input data. It has evolved into a prominent task in the realm of computer vision. Existing methods can be categorized into two main types: regression-based methods [4,11,51,60] and heatmap-based methods [29, 35, 49, 61]. Regression-based methods exhibit high efficiency for implementation on applications. However, these approaches inherently output singular 2D coordinates for each keypoint, neglecting to consider the surrounding area of the keypoint. To address this limitation, heatmap-based approaches have been introduced to localize keypoints through probabilistic heatmaps rather than fixed coordinates. In this paper, our method is based on heatmaps.

Category-Agnostic Pose Estimation aims to develop a pose estimation model that can detect the pose of various object categories, given few-shot examples as reference. Previous methods [5,28,45,46,57] typically train the model on a set of predefined base categories, and then adapt it on the given few-shot examples of unseen categories. In this work, we handle this task from a novel perspective that harnesses the knowledge within the off-the-shelf Stable Diffusion

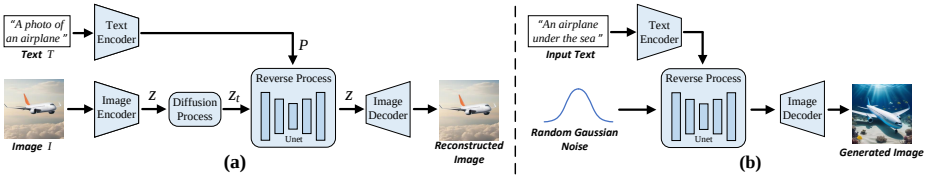

Fig. 2: **(a)** Overall architecture of Stable Diffusion. **(b)** The architecture of Stable Diffusion for text-to-image generation.

by leveraging the cross-attention mechanism to find correspondences between example images and test images, without requiring training on base categories.

Text-to-Image Diffusion Models [31, 36, 40, 43] are generation models that employ diffusion techniques. They are primarily utilized for generating detailed images based on textual descriptions. Recently, text-to-image diffusion models have made striking advancements across a spectrum of domains such as image inpainting [55], image editing [18], 3D scene understanding [26], semantic segmentation [20], visual tracking [62], object counting [16], and various interdisciplinary applications [30]. Their versatility and efficacy are reshaping the landscape of these fields. In this paper, we are the first to study the harness of text-to-image diffusion models for category-agnostic pose estimation.

3 Preliminary

In this paper, we propose a novel framework PPM to harness the knowledge in text-to-image diffusion models for CAPE. Our PPM has good extensibility, which can be applied to various text-to-image diffusion models. For clarity, in the main paper, we use Stable Diffusion [40] as an example to illustrate our framework. Below, we briefly review Stable Diffusion to facilitate a better understanding of our method.

Architecture. As shown in Fig. 2 (a), Stable Diffusion mainly consists of five parts: image encoder, image decoder, text encoder, diffusion process, and reverse process. Given an image I, the image is first mapped into a latent representation Z through an image encoder. Simultaneously, a text prompt T is transformed into an embedding P using a text encoder. The diffusion process is then applied to Z, which introduces noise over t steps to produce a noisy representation Z_t. Then, a reverse process incorporating a U-Net [41] takes in Z_t and P to denoise Z_t back to the clean Z. Finally, the image decoder reconstructs the image I from Z. After training on a vast amount of image-text pairs, the image encoder and the diffusion process are discarded, and the remaining parts can be regarded as a text-to-image generation model, as illustrated in Fig. 2 (b). During inference, we can input a random Gaussian noise into the generation model to generate a text-relevant image. This works because during training, the noisy Z_t can be very close to the pure Gaussian noise after many steps of the diffusion process.

Diffusion Training. As for training the diffusion model, the image encoder, the image decoder, and the text encoder are off-the-shelf frozen modules. The diffusion process does not involve learnable parameters. Therefore, it solely focuses on training the UNet (in the reverse process) to generate text-relevant visual content by denoising. In the diffusion process, noise ϵ is added to the latent image representation Z, resulting in Z_t. The training objective is to predict the added noise ϵ from Z_t, conditioned on P. This is formulated as:

$$L_{\text{dm}} = \mathbb{E}_{\mathcal{E}_i(I),\mathcal{E}_t(T),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(Z_t, P, t\right)\|_2^2\right], \tag{1}$$

where $\mathcal{E}_i$ is the image encoder, $\mathcal{E}_t$ represents the text encoder, t denotes the time step in the diffusion process, and ϵ_θ is the UNet in the reverse process. Through this training approach, the connection between textual information P and image information Z_t is established within the UNet.

4 Method

Problem Definition. This paper aims to address the task of Class-Agnostic Pose Estimation (CAPE). Unlike most pose estimation tasks that predict keypoints for a single known (seen) category, CAPE requires the model to efficiently generalize to novel categories, which have not only different appearances, but also varying numbers of keypoints. In CAPE, during training, we are given extensively labeled data of some predefined categories (called base categories). Here, we denote the data of base categories as D_{base}. In D_{base}, we use I_{base} and M_{base} to represent the image and the corresponding keypoint localization label, respectively. Following [57], we use the keypoint labels in heatmap format. For each image I_{base}, its corresponding heatmap label M_{base} contains multiple sub-labels, i.e., $M_{\text{base}} = \{M^1_{\text{base}}, M^2_{\text{base}}, ..., M^n_{\text{base}}, ..., M^N_{\text{base}}\}$ where each sub-label represents the heatmap of one keypoint, and N denotes the number of keypoints in the image. Previous methods [5, 28, 46, 57] generally train the model on D_{base} and then test the model on novel categories to validate its generalization capacity. Before testing, for each unseen category, the model is provided with one or a few labeled examples, denoted as D_{exm}, for few-shot learning. In D_{exm}, we use I_{exm} and M_{exm} to denote the example image and its corresponding label. Similarly, M_{exm} also contain multiple sub-labels for all keypoints, i,e, $M_{\text{exm}} = \{M^1_{\text{exm}}, M^2_{\text{exm}}, ..., M^n_{\text{exm}}, ..., M^N_{\text{exm}}\}$. During testing, based on the given few-shot examples D_{exm}, the model is required to detect the corresponding keypoints of the same category for the unlabeled test images I_{test}. Notably, since different object categories have varying numbers of keypoints, N may change accordingly based on the category. However, within the same category, all samples including example images and test images share the same number of keypoints.

In this paper, we propose a novel PPM framework that leverages the off-the-shelf Stable Diffusion to build the spatial (keypoint) correspondences between example images I_{exm} and text images I_{test}, thus addressing the CAPE task without the need of training on carefully prepared base categories D_{base}. Below, we present our PPM in detail.

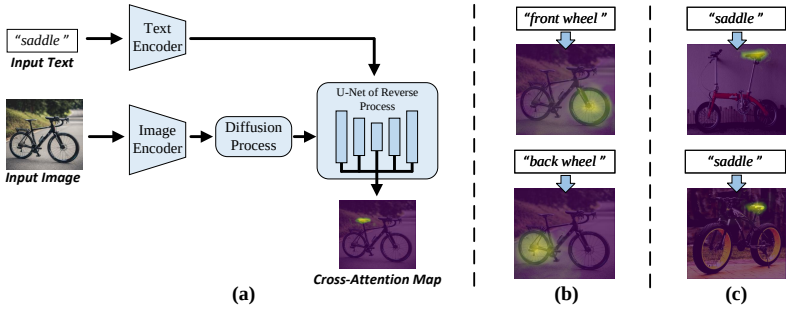

Fig. 3: **(a)** The architecture of Stable Diffusion for extracting cross-attention maps, which highlight the area that corresponds semantically to the input text. **(b)** With the same image, using different texts describing parts of the object, the attention maps highlight corresponding different areas. **(c)** With the same text "saddle", the cross-attention highlights relevant regions in different images.

4.1 Prompt Pose Matching (PPM)

Here, we propose a novel framework, Prompt Pose Matching (PPM), to effectively handle the CAPE task by learning only on the given few-shot examples D_{exm}. After learning on D_{exm}, our PPM can be directly used for inference on I_{test}, without training on D_{base}.

Inspired by the observation [20] that the off-the-shelf Stable Diffusion encompasses a wealth of knowledge that should be useful for understanding the object semantics in images, our PPM is designed to address CAPE by leveraging such Stable Diffusion embedded knowledge. To do so, we explore the cross-attention map in Stable Diffusion as a handle. We observe that the cross-attention map extracted from the U-Net of Stable Diffusion effectively captures the correspondence between the image and text. For example, as shown in Fig. 3 (a), when feeding an image of a bicycle into Stable Diffusion and providing the input text "saddle", the cross-attention map extracted from the U-Net highlights the corresponding saddle region, which occurs in a semantically consistent manner. Even when different images are fed into Stable Diffusion, if we use the same text prompt (e.g., "saddle"), the resulting cross-attention maps will consistently highlight corresponding regions, as shown in Fig. 3 (c). In light of this, given few-shot examples D_{exm} of an unseen category, if we can identify the text prompt T corresponding to a particular keypoint of this category, the diffusion model could also use this text prompt T to find the corresponding keypoints in test images I_{test} via cross-attention maps. However, since the actual text prompts of keypoints can be hard to obtain, we turn to learn a pseudo prompt P in the embedding space, aiming to properly represent the semantics of the particular keypoint. Based on this idea, our PPM framework is designed to handle the CAPE task via prompt optimization.

Overview. As illustrated in Fig. 4, our PPM involves two stages. In the first stage (prompt optimization), for the n-th keypoint in the given example image

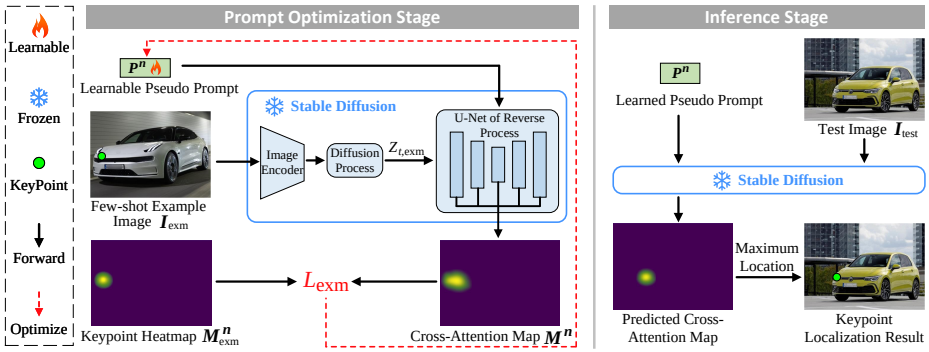

Fig. 4: Overview of our Prompt Pose Matching (PPM) framework.

I_{exm}, we seek to learn a pseudo prompt P^n that properly represents the semantics of this keypoint, by using the keypoint heatmap label M^n_{exm} to constrain the cross-attention map extracted from the U-Net of Stable Diffusion. For image I_{exm} with N keypoints, we learn N pseudo prompts $\{P^1, P^2, ..., P^n, ..., P^N\}$ respectively. In the second stage (inference), the N learned pseudo prompts from the first stage are kept fixed. With the help of these N learned prompts, N cross-attention maps are computed for the test image I_{test}. In each computed cross-attention map, the location with the highest attention value indicates the corresponding keypoint location of the test image I_{test}. In this way, we use the learned pseudo prompts to bridge the keypoint correspondences between example images I_{exm} and text images I_{test}, which subtly harnesses Stable Diffusion to address CAPE. Next, we introduce the two stages of our PPM.

Prompt Optimization Stage. In this stage, we aim to learn pseudo prompts to represent the semantics of the keypoints from few-shot examples D_{exm}. Taking the learning for the n-th keypoint as an example, at the beginning, the pseudo prompt $P^n \in \mathbb{R}^{1\times1\times c}$ is randomly initialized, where c is the channel number of the embedding space. Also, we are given a test-time example image $I_{\text{exm}} \in \mathbb{R}^{H\times W\times 3}$ and the n-th keypoint heatmap label $M^n_{\text{exm}} \in \mathbb{R}^{H\times W\times 1}$. As shown in Fig. 4 (left), we input the prompt P^n along with the example image I_{exm} into the stable diffusion to obtain the cross-attention map M^n. Specifically, the image encoder and the diffusion process of Stable Diffusion convert the image I_{exm} into the noisy latent representation $Z_{t,\text{exm}}$. Then, $Z_{t,\text{exm}}$ and P^n are fed into the U-Net of the reverse process. At each layer of the U-Net, a cross-attention map can be obtained, which calculates the similarity (correlation) between $Z_{t,\text{exm}}$ and P^n. More specifically, for the n-th keypoint, the cross-attention map at the l-th layer in the U-Net can be obtained as:

$$M^n_l = Sigmoid\big(Avg(\frac{F_{Z_{t,\text{exm}}} * F^{\top}_{P^n}}{\sqrt{d_l}})\big), \tag{2}$$

where $F_{Z_{t,\text{exm}}} \in \mathbb{R}^{H_l\times W_l\times d_l}$ is the visual feature map, and $F_{P^n} \in \mathbb{R}^{1\times1\times d_l}$ is the prompt feature; $F_{Z_{t,\text{exm}}}$ and F_{P^n} are obtained from $Z_{t,\text{exm}}$ and P^n, respectively,

via $F_{Z_{t,\text{exm}}} = \phi_l(Z_{t,\text{exm}})$ and $F_{P^n} = \psi_l(P^n)$, where ϕ_l and ψ_l denote the linear mappings in the attention module at the l-th layer of the U-Net; H_l, W_l, and d_l respectively represent the height, width, and channel number of the visual feature map at the l-th layer; Avg denotes the average along the channel dimension; $Sigmiod$ is the normalization that transforms the values into $[0, 1]$. In Eqn. 2, the prompt feature F_{P^n} is multiplied with each pixel-level visual feature of the feature map $F_{Z_{t,\text{exm}}}$, and thus if the prompt feature and visual feature are semantically similar (correlated), the attention value at the corresponding pixel will be high, thereby highlighting the prompt-related regions. We denote the l-th layer cross-attention map as M_l, where $M_l \in \mathbb{R}^{H_l \times W_l \times 1}$.

To incorporate cross-attention maps from different layers, as shown in Fig. 4 (left), we average cross-attention maps across the U-Net layers, to obtain:

$$M^n = \frac{1}{L} \sum_{l=1}^{L} M_l^n, \tag{3}$$

where L denotes the number of layers in U-Net. Note that when averaging, we first use the bilinear interpolation to upsample all the cross-attention maps into the original image scale $H \times W$. Thus, we can obtain $M^n \in \mathbb{R}^{H \times W \times 1}$.

After obtaining the cross-attention map M^n for the n-th keypoint of the example image, given the keypoint heatmap label M^n_{exm} of this image, we calculate the L2 loss between M^n and M^n_{exm}, and use the loss to optimize the pseudo prompt P^n, ensuring that P^n can drive Stable Diffusion to highlight the specific keypoint region. The optimization loss can be formulated as:

$$L_{\text{exm}} = \|M^n - M^n_{\text{exm}}\|^2 + L_{\text{dm}}, \tag{4}$$

where $\|\cdot\|^2$ denotes the L_2 distance. The second items L_{dm} is the diffusion model's loss function (Eq. 1). We use this loss to keep the learned prompt understandable by Stable Diffusion. After optimization, we can obtain a learned pseudo prompt P^n that represents the semantics of the n-th keypoint. By learning from $n = 1$ to $n = N$, we can obtain N learned pseudo-prompts. Not that the loss is calculated across all the given few-shot examples by averaging. That is, for a novel category (N keypoints) with multiple labeled examples (i.e., more than 1-shot), we still obtain N learned pseudo-prompts by averaging across the examples.

Next, we describe how to use the learned prompt P^n to conduct keypoint localization for the test image I_{test} (i.e., the inference stage).

Inference Stage. To build the image-to-image correspondences for CAPE, we use the prompt P^n learned from example images, to find the corresponding keypoint in each test image. As shown in Fig. 4 (right), given a test image I_{test}, we feed P^n and I_{test} into Stable Diffusion to compute the cross-attention map. The highlighted region in the output cross-attention map denotes the detected corresponding keypoint region. As P^n is learned from the keypoint regions in the example images using the knowledge of Stable Diffusion, the P^n-activated keypoint regions in the test images share the same semantic meaning and structural characteristics as those in the example images. Finally, we locate the keypoint by

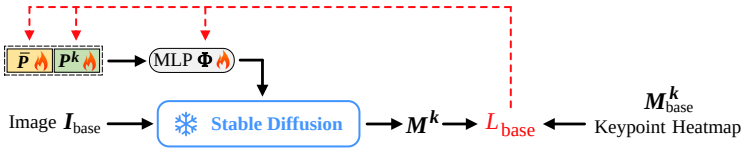

Fig. 5: Illustration of our PPM with Category-shared Prompt Training (CPT). $\bar{P}$ and P^k denote the category-shared prompt and category-specific prompt, respectively. Φ denotes a light-weight MLP.

finding the position with maximum value in the cross-attention map, as shown in Fig. 4 (right). In this way, based on all learned prompts $\{P^1, P^2, ..., P^n, ..., P^N\}$, we can estimate the location of all keypoints for test images of the same category.

4.2 Category-shared Prompt Training (CPT)

Though keypoints of different categories show various characteristics, there can be some common knowledge shared by keypoints of different categories, which could help for keypoint localization (e.g., keypoints tend to be located at "important" locations, like center points, corners, edges, etc.). In light of this, we propose to capture common knowledge by additionally learning a category-shared prompt. Specifically, we propose a Category-shared Prompt Training (CPT) scheme, to enable our framework to learn one category-shared prompt $\bar{P}$ from data of base categories D_{base} provided in the CAPE task. Our goal is to train this category-shared prompt $\bar{P}$ on each keypoint of every category from base categories, thus enabling $\bar{P}$ to acquire category-agnostic knowledge. This knowledge can equally benefit pose estimation for novel categories.

To disentangle the shared knowledge and the specific knowledge during the training on base categories, we slightly modify the PPM framework. We assume that across all categories in the base-category data, there are a total of K keypoints. To represent these K keypoints precisely, we require K category-specific prompts, denoted as $\{P^1, P^2, ..., P^k, ..., P^K\}$. As shown in Fig. 5, before sending it into the Stable Diffusion, we concatenate the learnable category-shared prompt $\bar{P}$ with the learnable category-specific prompt P^k, and then pass the combined prompt through a light-weight MLP [6] (two layers) to reduce the channel dimension, aiming to ensure compatibility with the Stable Diffusion. Specifically, given I_{base} (the image from base categories) and M^k_{base} (the corresponding heatmap label of the k-th keypoint), the training loss L can be calculated as:

$$M^k = f_{SD}\big(I_{\text{base}}, \Phi([\bar{P}, P^k])\big), \tag{5}$$

$$L_{\text{base}} = \left\| M^k - M^k_{\text{base}} \right\|^2 + L_{\text{dm}}, \tag{6}$$

where M^k is the output cross-attention map; f_{SD} denotes the Stable Diffusion which treats I_{base} and $\Phi([\bar{P}, P^k])$ as two inputs; Φ denotes the MLP; $\bar{P}$ and P^k

represent the category-shared prompt and the category-specific prompt respectively. As shown in Fig. 5, we use the loss in Eqn. 6 to optimize the category-shared prompt $\bar{P}$, category-specific prompt P^k, and MLP Φ, where the Stable Diffusion is kept frozen. During this training process, the (one) category-shared prompt $\bar{P}$ and the MLP Φ are trained on all keypoints of the base categories, while the category-specific prompt P^k is independently trained for each keypoint of each base category. Thus, after training on K base-category keypoints, we will get K learned category-specific prompts $\{P^1, P^2, ..., P^K\}$ and only one learned category-shared prompt $\bar{P}$. In this disentangled learning way, $\bar{P}$ learns the common knowledge of keypoint localization, P^k learns the specific knowledge, and Φ learns how to merge the common knowledge and specific knowledge. As each learned P^k is specific to the keypoint of base categories, we discard all the learned P^k, but introduce the learned $\bar{P}$ and the learned Φ into our PPM framework to help to enhance the category-specific pseudo prompt P^n learned from few-show examples of unseen categories.

4.3 Training and Testing

As for PPM *without* CPT, we do not use data of based categories. Given few-shot examples of an unseen category, for each keypoint of this category, we randomly initialize a pseudo prompt P^n and use the labeled examples to optimize P^n. Then during testing, we use the learned P^n to locate the corresponding keypoint for test images.

As for PPM *with* CPT, it involves training on base categories. During training on base categories, we use labeled data of base categories to train a category-shared prompt $\bar{P}$ and a MLP Φ. Then, we initialize the pseudo prompt P^n for learning on few-shot examples. We concatenate the learnable P^n with the learned (frozen) category-shared $\bar{P}$ using the learned (frozen) MLP Φ, to enhance the prompt representation. We follow the same procedures of PPM *without* CPT to optimize P^n on few-shot examples. Finally, during testing, we still concatenate the learned P^n with the learned $\bar{P}$ using the learned MLP Φ, to produce the enhanced prompt for inference on test images.

5 Experiments

Dataset and Metric. The majority of pose estimation datasets are not suitable for the CAPE task, since they predominantly consist of objects belonging to a single category. Following prior CAPE research [57], we evaluate our approach on the benchmark MP-100 [57], which is composed of many popular 2D pose estimation datasets, including COCO [25], 300W [42], AFLW [22], OneHand10K [52], DeepFasion2 [9], MacaquePose [24], Vinegar Fly [33], Desert Locust [12], CUB-200 [53], CarFusion [37], AnimalWeb [19], and Keypoint-5 [54]. In total, MP-100 covers eight super-categories (i.e., *human hand*, *human face*, *human body*, *animal face*, *animal body*, *clothes*, *furniture*, and *vehicle*) and 100 sub-categories (e.g.,

bus, *sofa*, *bed*, and *skirt*, etc), providing a very comprehensive evaluation platform. The MP-100 benchmark is organized into training, validation, and testing sets, comprising categories without overlapping, thus facilitating the evaluation of models for CAPE. On this benchmark, we use the standard metric of PCK (Probability of Correct Keypoint) [59] with a threshold of 0.2 to assess the algorithm's performance, following previous work [45, 57].

Parameter Setting. The training is conducted on two NVIDIA RTX 3090 GPUs with a batch size of four in each GPU. During training in CPT, we adopt the Adam optimizer [21] with a learning rate of $5e-4$. During the test-time optimization on few-shot examples, the learning rate is set to $2e-3$. In the diffusion process of Stable Diffusion, we use $t = 10$ to obtain z^t.

Experiment Settings. In CAPE, there are two generalization settings: (1) Cross Super-Category Generalization and (2) Cross Sub-Category Generalization. In Cross Super-Category Generalization, the generalization performance is evaluated based on 8 super-categories, where 6 for train, 1 for val, and 1 for test. In Cross Sub-Category Generalization, the evaluation is conducted based on 100 sub-categories, where 70 for train, 10 for val, and 20 for test. Besides, there are also two few-shot settings: (1) 1-shot and (2) 5-shot, which represent the number of few-shot examples. As existing methods [5, 28, 45, 46, 57] conduct experiments under three scenarios: Cross Super-Category Generalization (1-shot), Cross Sub-Category Generalization (1-shot), and Cross Sub-Category Generalization (5-shot), for fair comparisons, we follow them to conduct experiments under the same 3 scenarios.

Table 1: Results in the setting of Cross Super-Category Generalization (1-shot). ✓ and ✗ denote use or not use of base categories.

Method	Base Cate.	Human Body	Human Face	Vehicle	Furniture	Mean
ProtoNet [46]	✓	37.61	57.80	28.35	42.64	41.60
MAML [5]	✓	51.93	25.72	17.68	20.09	23.08
Fine-tune [28]	✓	52.11	25.53	17.46	20.76	28.97
POMNet [57]	✓	73.82	79.63	34.92	47.27	58.91
CapeFormer [45]	✓	83.44	80.96	45.40	52.49	65.57
PPM	✗	84.03	81.52	50.59	58.61	68.68
PPM + CPT	✓	**85.13**	**82.44**	**52.08**	**60.59**	**70.06**

5.1 Quantitative Results

Cross Super-Category Generalization Results. In MP-100 dataset, there are eight super-categories. We follow [45, 57] to undertake a cross super-category pose estimation evaluation on the MP-100 benchmark. Specifically, the model is trained and validated on six and one super-categories, and performance is evaluated on the remaining one super-category. For fair comparisons, we follow previous work [45, 57] to evaluate the generalization to *human body*, *human face*, *vehicle*, and *furniture*, respectively.

As shown in Tab. 1, the performance of our proposed PPM, even with no base categories, surpasses that of previous methods trained on base categories, clearly demonstrating the effectiveness of our framework. We can see that previous methods get suboptimal performances. This could be attributed to the significant divergence between training categories and the testing one, which hinders the generalization of methods that highly rely on knowledge from the training base categories. In contrast, our approach leverages the wealth of knowledge within the Stable Diffusion model to mitigate overfitting to the training categories, thus achieving a significant performance improvement over prior methods (e.g., on *furniture*, we outperform [45] by 8.1%). It is noteworthy that even using PPM only, our method can significantly outperform existing methods, and with the addition of CPT, the performance of our method is further boosted, which demonstrates the effect of our method.

Table 2: Results in the setting of Cross Sub-Category Generalization.

Method	Base Cate.	1-shot						5-shot					
		Split1	Split2	Split3	Split4	Split5	Mean	Split1	Split2	Split3	Split4	Split5	Mean
ProtoNet [46]	✓	46.05	40.84	49.13	43.34	44.54	44.78	60.31	53.51	61.92	58.44	58.61	58.56
MAML [5]	✓	68.14	54.72	64.19	63.24	57.20	61.50	70.03	55.98	63.21	64.79	58.47	62.50
Fine-tune [28]	✓	70.60	57.04	66.06	65.00	59.20	63.58	71.67	57.84	66.76	66.53	60.24	64.61
POMNet [57]	✓	84.23	78.25	78.17	78.68	79.17	79.70	84.72	79.61	78.00	80.38	80.85	80.71
CapeFormer [45]	✓	89.45	84.88	83.59	83.53	85.09	85.31	91.94	88.92	89.40	88.01	88.25	89.30
PPM	✗	89.82	85.63	83.72	83.90	86.03	85.82	92.31	89.43	89.84	89.97	89.30	90.17
PPM+CPT	✓	**91.03**	**88.06**	**84.48**	**86.73**	**87.40**	**87.54**	**93.64**	**92.71**	**91.76**	**92.85**	**91.94**	**92.58**

Cross Sub-Category Generalization Results. To further assess generalization capability to unseen sub-categories, following [45, 57], we evaluate our framework against previous methods under two few-shot settings: 1-shot and 5-shot, which are shown in Tab. 2. The evaluation is conducted on five different train/val/test category splits for a comprehensive evaluation. In both 1-shot and 5-shot settings, we show the results of each split and the mean results averaged over all the five splits. From Tab. 2, we can see: (1) in the absence of CPT, our PPM, even using no training data of base categories, can outperform previous methods trained on base categories across all splits and few-shot settings. (2) When using PPM equipped with CPT, our method achieves further enhanced performance and significantly surpasses existing methods, which demonstrates the effectiveness of CPT.

5.2 Qualitative Results

In Fig. 6, we qualitatively compare the generalization ability of our method with existing approaches. We can observe that when the few-shot example largely differs from the test images in terms of viewpoint or appearance, our method still performs robustly and generates predictions very close to the ground-truth labels. To demonstrate our CPT can effectively enhance our PPM's keypoint identification capacity, we further visualize cross-attention maps of the test set.

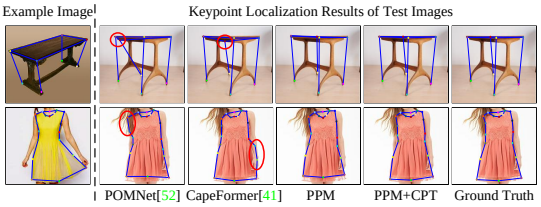

Fig. 6: Visual comparison with state-of-the-art methods. Red circles show failure keypoints.

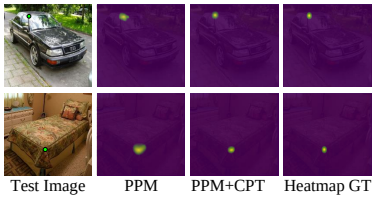

Fig. 7: Visualization of attention maps in our method.

As shown in Fig. 7, CPT can drive Stable Diffusion to generate cross-attention maps with finer and preciser keypoint regions. This indicates that, compared to only use category-specific prompt, the addition of category-shared prompt provided by CPT facilitates the Stable Diffusion to better locate the keypoints.

5.3 Ablation Study

Effect of incorporating multi-layer attention maps. In our PPM framework, we integrate cross-attention maps across different layers. In this study, we explore the impact of incorporating multi-layer cross-attention maps on the overall performance of the framework. For comparison, we evaluate the performance of cross-attention maps at each single layer. Because of the limited space, in Tab. 3, we only present the best result among all single-layer cross-attention maps. We can see that the integration of cross-attention maps across different layers shows an obvious performance advantage, even compared to the best-performing single-layer cross-attention map. More ablation studies can be seen in *Supplementary*.

Table 3: Ablation on incorporation of multi- layer attention maps.

Method	Cross Super-Cate. (1-shot)	Cross Sub-Cate. (1-shot)	Cross Sub-Cate. (5-shot)
Ours (Single Layer)	68.81	86.49	91.37
Ours (Multi Layer)	**70.06**	**87.54**	**92.58**

6 Conclusion

In this paper, we introduce a novel PPM framework to address CAPE by harnessing the knowledge in text-to-image diffusion models, without requiring base categories. We further propose a CPT scheme to enhance the keypoint identification capability of PPM. Experiments show that our method significantly outperforms existing state-of-the-art methods.

Acknowledgements This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2022-01-027[T]) and the Ministry of Education, Singapore, under the AcRF Tier 2 Projects (MOE-T2EP20222-0009 and MOE-T2EP20123-0014).

References

1. Bulat, A., Tzimiropoulos, G.: Human pose estimation via convolutional part heatmap regression. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14. pp. 717–732. Springer (2016) 4
2. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE international conference on computer vision. pp. 1021–1030 (2017) 1, 4
3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7291–7299 (2017) 1
4. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5933–5942 (2019) 4
5. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International conference on machine learning. pp. 1126–1135. PMLR (2017) 2, 4, 6, 12, 13
6. Gardner, M.W., Dorling, S.: Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric environment **32**(14-15), 2627–2636 (1998) 10
7. Ge, L., Cai, Y., Weng, J., Yuan, J.: Hand pointnet: 3d hand pose estimation using point sets. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8417–8426 (2018) 4
8. Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., Cai, J., Yuan, J.: 3d hand shape and pose estimation from a single rgb image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10833–10842 (2019) 4
9. 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) 11
10. Gilroy, S., Glavin, M., Jones, E., Mullins, D.: Pedestrian occlusion level classification using keypoint detection and 2d body surface area estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3833–3839 (2021) 1
11. Gong, J., Fan, Z., Ke, Q., Rahmani, H., Liu, J.: Meta agent teaming active learning for pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11079–11089 (2022) 4
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) 11
13. Guleryuz, O.G., Kaeser-Chen, C.: Fast lifting for 3d hand pose estimation in ar/vr applications. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 106–110 (2018) 1

14. Huang, M.H., Foo, L.G., Liu, J.: Learning to unlearn for robust machine unlearning. In: European Conference on Computer Vision. Springer (2024) 1
15. Huang, S., Qi, S., Xiao, Y., Zhu, Y., Wu, Y.N., Zhu, S.C.: Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation. Advances in Neural Information Processing Systems **31** (2018) 1
16. Hui, X., Wu, Q., Rahmani, H., Liu, J.: Class-agnostic object counting with text-to-image diffusion model. In: European Conference on Computer Vision. Springer (2024) 5
17. Iftikhar, S., Zhang, Z., Asim, M., Muthanna, A., Koucheryavy, A., Abd El-Latif, A.A.: Deep learning-based pedestrian detection in autonomous vehicles: Substantial issues and challenges. Electronics **11**(21), 3551 (2022) 1
18. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023) 5
19. 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) 11
20. Khani, A., Taghanaki, S.A., Sanghi, A., Amiri, A.M., Hamarneh, G.: Slime: Segment like me. arXiv preprint arXiv:2309.03179 (2023) 2, 3, 5, 7
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 12
22. 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) 11
23. Krauß, V., Boden, A., Oppermann, L., Reiners, R.: Current practices, challenges, and design implications for collaborative ar/vr application development. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–15 (2021) 1
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) 11
25. 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: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 1, 4, 11
26. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9298–9309 (2023) 5
27. Liu, Z., Chen, Z., Bai, J., Li, S., Lian, S.: Facial pose estimation by deep learning from label distributions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019) 4
28. Nakamura, A., Harada, T.: Revisiting fine-tuning for few-shot learning. arXiv preprint arXiv:1910.00216 (2019) 2, 4, 6, 12, 13
29. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. pp. 483–499. Springer (2016) 4

30. Nguyen, L.X., Aung, P.S., Le, H.Q., Park, S.B., Hong, C.S.: A new chapter for medical image generation: The stable diffusion method. In: 2023 International Conference on Information Networking (ICOIN). pp. 483–486. IEEE (2023) 5
31. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In: Proceedings of the 39th International Conference on Machine Learning. vol. 162, pp. 16784–16804. PMLR (17–23 Jul 2022) 2, 5
32. Peng, H., Huang, H., Xu, L., Li, T., Liu, J., Rahmani, H., Ke, Q., Guo, Z., Wu, C., Li, R., et al.: The multi-modal video reasoning and analyzing competition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 806–813 (2021) 1
33. 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) 11
34. Probst, T., Fossati, A., Van Gool, L.: Combining human body shape and pose estimation for robust upper body tracking using a depth sensor. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14. pp. 285–301. Springer (2016) 1
35. Ramakrishna, V., Munoz, D., Hebert, M., Bagnell, J.A., Sheikh, Y.: Pose machines: Articulated pose estimation via inference machines. In: European Conference on Computer Vision. pp. 33–47. Springer (2014) 4
36. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 2, 5
37. 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) 11
38. Reddy, N.D., Vo, M., Narasimhan, S.G.: Occlusion-net: 2d/3d occluded keypoint localization using graph networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7326–7335 (2019) 4
39. Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine-grained visual descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016) 3
40. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 2, 5
41. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015) 5
42. 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) 11
43. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022) 2, 5

44. Sánchez, H.C., Martínez, A.H., Gonzalo, R.I., Parra, N.H., Alonso, I.P., Fernandez-Llorca, D.: Simple baseline for vehicle pose estimation: Experimental validation. IEEE Access **8**, 132539–132550 (2020) 4
45. 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) 2, 4, 12, 13
46. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in neural information processing systems **30** (2017) 2, 4, 6, 12, 13
47. 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) 1
48. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3476–3483 (2013) 4
49. Tang, W., Wu, Y.: Does learning specific features for related parts help human pose estimation? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1107–1116 (2019) 4
50. 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) 4
51. Wang, J., Long, X., Gao, Y., Ding, E., Wen, S.: Graph-pcnn: Two stage human pose estimation with graph pose refinement. In: European Conference on Computer Vision. pp. 492–508. Springer (2020) 4
52. 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) 11
53. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-ucsd birds 200 (2010) 11
54. Wu, J., Xue, T., Lim, J.J., Tian, Y., Tenenbaum, J.B., Torralba, A., Freeman, W.T.: Single image 3d interpreter network. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14. pp. 365–382. Springer (2016) 11
55. Xie, S., Zhang, Z., Lin, Z., Hinz, T., Zhang, K.: Smartbrush: Text and shape guided object inpainting with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22428–22437 (2023) 5
56. Xu, L., Huang, M.H., Shang, X., Yuan, Z., Sun, Y., Liu, J.: Meta compositional referring expression segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19478–19487 (2023) 1
57. 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) 2, 4, 6, 11, 12, 13
58. Xu, W., Su, P.c., Sen-ching, S.C.: Human pose estimation using two rgb-d sensors. In: 2016 IEEE International Conference on Image Processing (ICIP). pp. 1279–1283. IEEE (2016) 1
59. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE transactions on pattern analysis and machine intelligence **35**(12), 2878–2890 (2012) 12

60. Zhang, D., Guo, G., Huang, D., Han, J.: Poseflow: A deep motion representation for understanding human behaviors in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6762–6770 (2018) 4
61. Zhang, J., Cai, Y., Yan, S., Feng, J., et al.: Direct multi-view multi-person 3d pose estimation. Advances in Neural Information Processing Systems **34**, 13153–13164 (2021) 4
62. Zhang, Z., Xu, L., Peng, D., Rahmani, H., Liu, J.: Diff-tracker: Text-to-image diffusion models are unsupervised trackers. In: European Conference on Computer Vision. Springer (2024) 5
63. Zhang, Z., Zhou, C., Tu, Z.: Distilling inter-class distance for semantic segmentation. arXiv preprint arXiv:2205.03650 (2022) 1
64. Zhao, C., Fu, C., Dolan, J.M., Wang, J.: L-shape fitting-based vehicle pose estimation and tracking using 3d-lidar. IEEE Transactions on Intelligent Vehicles **6**(4), 787–798 (2021) 4
65. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11656–11665 (2021) 1
66. Zimmermann, C., Brox, T.: Learning to estimate 3d hand pose from single rgb images. In: Proceedings of the IEEE international conference on computer vision. pp. 4903–4911 (2017) 1, 4