# Unleashing the Power of Text-to-Image Diffusion Models for Category-Agnostic Pose Estimation

Duo Peng, Zhengbo Zhang, Ping Hu, Qiuhong Ke, De Wen Soh, Mohammed Bennamoun, and Jun Liu

**Abstract**—Category-Agnostic Pose Estimation (CAPE) aims to detect keypoints of unseen object categories in a few-shot setting, where the scarcity of labeled data poses significant challenges to generalization. In this work, we propose Prompt Pose Matching (PPM), a novel framework that unleashes the power of off-the-shelf text-to-image diffusion models for CAPE. PPM learns pseudo prompts from 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. To provide prompts with representative initialization, we introduce a category-agnostic pre-training strategy to capture the foreground prior shared across categories and keypoints. To support the reliable prompt pre-training, we propose a Foreground-Aware Region Aggregation (FARA) module to provide robust and consistent supervision signal. Based on the foreground prior, a Foreground-Guided Attention Refinement (FGAR) module is further proposed to reinforce cross-attention responses for accurate keypoint localization. For efficiency, a Prompt Ensemble Inference (PEI) scheme enables joint keypoint prediction. Unlike previous methods that highly rely on base-category annotated data, our PPM framework can operate in a base-category-free setting while retaining strong performance. Code will be available at: https://github.com/DuoPeng-CVer/Prompt-Pose-Matching.

**Index Terms**—Category-Agnostic Pose Estimation, Diffusion Model, Text-to-Image Generation, Prompt Learning.

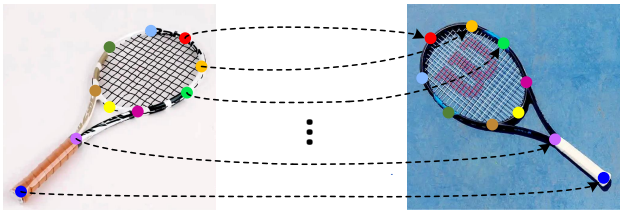
## 1 Introduction

Pose estimation aims to localize human-defined keypoints in images, with applications in AR/VR [1], [2], autonomous driving [3], [4], and health care [5], [6]. Existing methods [7], [8] achieve strong performance but are typically trained for a single category, limiting their generalization to diverse unseen categories in real-world scenarios.

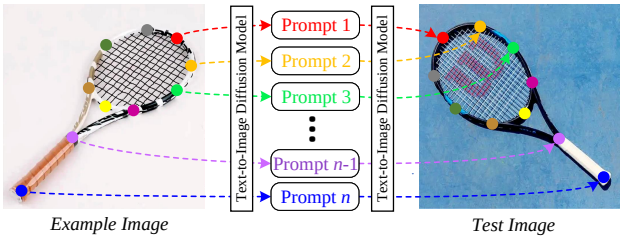
Recent research [9] has introduced Category-Agnostic Pose Estimation (CAPE) to address this generalization challenge. In CAPE, the model is provided with one or a few examples of an unseen category before testing. After adapting to these few-shot examples, the model is expected to localize keypoints in test images that share the same semantic and structural meanings, thereby achieving pose estimation in a category-agnostic manner with few-shot supervision.

Given the limited information from few-shot examples, existing CAPE methods [9], [10], [11], [12], [13] typically rely on external knowledge, i.e., training on predefined base categories and then adapting to unseen ones. While effective, this requires extensive data collection and labeling, and models still struggle to generalize from only a few examples. Prior studies [9], [13] show that even with large-scale training, performance drops significantly on unseen categories (e.g., *furniture*, *vehicle*) that differ from the base categories (e.g., *face*, *human body*). These challenges highlight the need for a more data-efficient and generalizable approach to CAPE.

Driven by recent advances in cross-modal representation




**(a)** Correspondences in CAPE task




**(b)** Correspondences built by our PPM

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, and enable us to harness the knowledge of the text-to-image diffusion model. Note that in both (a) and (b), there are multiple correspondences. Here we only show a few of them for clarity.

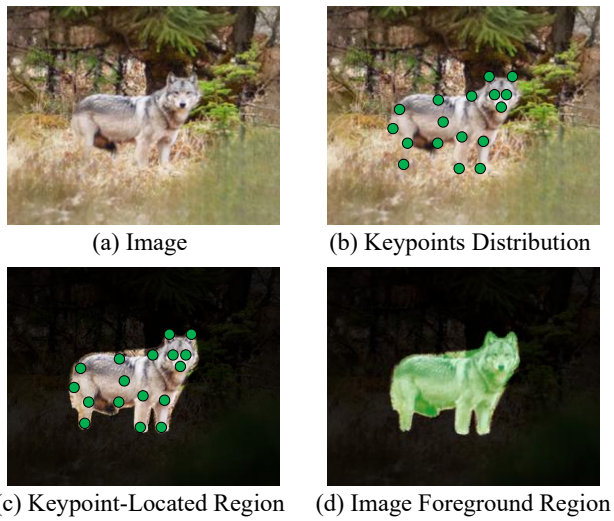
*This research/project is supported by the Ministry of Education, Singapore under its Research Centre of Excellence award to the Institute for Digital Molecular Analytics & Science, NTU (IDMxS, grant: EDUNC-33-18-279-V12), the National Natural Science Foundation of China (Grant No: 62476048), and the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2022-01-027).*

- *(Corresponding author: Jun Liu.)*
- *D. Peng is with the Institute for Digital Molecular Analytics and Science, Nanyang Technological University, Singapore. E-mail: duo.peng@ntu.edu.sg.*
- *Z. Zhang, and De Wen Soh are with the Singapore University of Technology and Design, Singapore. E-mail: zhengbo_zhang@mymail.sutd.edu.sg, and dewen_soh@sutd.edu.sg.*
- *P. Hu is with University of Electronic Science and Technology of China, China, and VinUniversity, Vietnam. E-mail: chinahuping@gmail.com.*
- *Q. Ke is with the Monash University, Australia. E-mail: Qiuhong.Ke@monash.edu.*
- *M. Bennamoun is with the University of Western Australia, Australia. E-mail: mohammed.bennamoun@uwa.edu.au.*
- *J. Liu is with the Lancaster University, United Kingdom. E-mail: j.liu81@lancaster.ac.uk.*

learning [14], [15], [16], [17], text-to-image (T2I) diffusion models [18], [19], [20], like Stable Diffusion [21], have recently gained attention for their remarkable ability to generate photo-realistic images based on user-specified prompts. Given that existing T2I 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 T2I 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. Inspired by this insight, in this work, we propose to harness the power of T2I diffusion models to address the CAPE task that involves spatial composition reasoning, by learning from only few-shot examples of unseen categories. The challenge, however, lies in that T2I diffusion models are designed for image generation, not keypoint localization, making it non-trivial to directly exploit their knowledge.

In response to the challenge mentioned above, this paper handles the CAPE task from a new perspective, which regards this task as establishing spatial correspondences (mappings) between example images and test images, as shown in Fig. 1 (a). This perspective inspires us a feasible way to leverage T2I diffusion models to address CAPE, raising a question: *Given that the T2I 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 T2I diffusion models. In particular, we focus on the commonly-used cross-attention mechanism in T2I diffusion models, which facilities the visual-textual interactions in the text-to-image generation process. We observe that the cross-attention maps extracted from T2I diffusion models help to highlight the text-related regions of the images. For example, 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, 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.

Building on this insight, we aim to extend the capabilities of T2I 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 T2I 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 [22], since many keypoints (e.g., keypoints in Fig. 1) can be difficult to describe in texts even for human beings.



(a) Image (b) Keypoints Distribution (c) Keypoint-Located Region (d) Image Foreground Region

Fig. 2. **(a)** An image sample in CAPE. **(b)** The spatial distribution of all labeled keypoints. **(c)** The region where the keypoints are located. **(d)** The foreground region of the image, highlighted in green. From (c) and (d), we can see that all keypoints lie within the foreground region.

In this work, we introduce a novel framework called **Prompt Pose Matching (PPM)** to address this challenge. The core mechanism of PPM is a few-shot prompt adaptation strategy, which identifies appropriate 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 effectively leverage its inherent knowledge to localize keypoints for unseen categories.

With the above few-shot prompt adaptation (core mechanism of PPM) in place, we further develop several auxiliary components to ensure the PPM's reliability in the CAPE setting. The following three components (one for prompt initialization, one for attention refinement, and one for ensemble inference) serve as supporting designs that make the PPM framework more stable, accurate, and efficient.

As for prompt initialization, a key observation across both base and novel categories is that keypoints consistently lie on the foreground object region (see Fig. 2). Motivated by this, we propose to learn a category-agnostic foreground prompt that highlights object regions across diverse categories. This is because initializing keypoint-specific prompts from this foreground-aware prior offers a much stronger starting point than random initialization, which is crucial under few-shot supervision. Since CAPE provides only keypoint annotations, we propose a **Foreground-Aware Region Aggregation (FARA)** module to derive pseudo foreground labels from keypoints by expanding them to visually coherent neighboring regions using diffusion-model affinity cues. These pseudo

labels enable effective training of a generalizable foreground prompt without extra supervision.

As for attention refinement, the learned foreground prompt also serves as a prior for sharpening keypoint cross-attention maps. Our proposed **Foreground-Guided Attention Refinement (FGAR)** module suppresses irrelevant activations and strengthens attention responses around keypoint-relevant regions, alleviating the inaccuracy of raw cross-attention maps.

As for ensemble inference, we propose a **Prompt Ensemble Inference (PEI)** scheme to improve efficiency. PEI aggregates multiple prompts and predicts several keypoints in a single forward pass, avoiding the one-keypoint-per-inference limitation common in prior CAPE methods.

While the proposed PPM framework is introduced in the context of training on base categories to obtain a generalizable foreground prior, we further extend it to a variant that requires no such training. Remarkably, all key components—FARA, FGAR, and PEI—remain applicable and effective even without access to base-category data.

To provide a clear overview of the entire framework, we summarize the full PPM pipeline as follows. First, we use FARA module (Sec. 3.1) to learn a category-agnostic foreground prompt that provides a strong initialization. Building on this initialized prompt, we then perform few-shot adaptation (Sec. 3.2) to obtain keypoint-specific pseudo prompts, each corresponding to one keypoint of novel categories. During inference on novel categories, we enhance cross-attention quality through a training-free FGAR module (Sec. 3.3), and finally use PEI scheme (Sec. 3.4) to predict multiple keypoints efficiently.

We summarize our contributions below:

- A novel framework, PPM, is introduced to exploit the power of T2I diffusion models for CAPE.
- To enable better adaptation to unseen categories, we propose to learn a class-agnostic foreground prompt as the initialization.
- We design FARA, a keypoint-seeded region aggregation module that generates pseudo foreground masks from sparse annotations.
- We introduce FGAR, a refinement module that distills foreground prior into cross-attention maps, enabling more accurate and focused keypoint localization.
- We develop PEI, a prompt ensemble scheme that enables efficient multi-keypoint localization.

This paper is an extension of our preliminary conference paper [23] (oral presentation) with several major improvements: **(1)** We replace the original CPT scheme with a new class-agnostic prompt pre-training strategy, which learns a foreground-aware initialization and provides a clearer, more effective starting point for test-time keypoint localization. **(2)** To better utilize limited keypoint annotations, we introduce the FARA module to derive pseudo foreground masks via local region aggregation. **(3)** We refine the attention mechanism with the FGAR module, which enhances cross-attention using the learned foreground prior. **(4)** We develop the PEI scheme for multi-keypoint inference, substantially improving efficiency. **(5)** Beyond MP-100, we evaluate our method on the more challenging SPair-71k dataset. **(6)** We further demonstrate the extendability of PPM on a distinct task: Unsupervised Dense Correspondences (UDC).

## 2 Related Work

### 2.1 Pose Estimation

Pose estimation endeavors to localize semantic or interest keypoints, such as human body parts [24], [25], [26], [27], [28], facial landmarks [29], [30], [31], hand keypoints [32], [33], [34], and vehicle poses [35], [36], [37] from input images. This task has become more important in computer vision, with applications spanning from augmented reality to robotics. Existing methods can be categorized into two main types: regression-based methods [38], [39], [40], [41] and heatmap-based methods [42], [43], [44], [45]. Regression-based methods [38], [39], [40], [41] are known for their high efficiency and speed, making them suitable for real-time applications. These approaches typically output a single set of 2D coordinates for each keypoint, which can streamline the pose estimation process. However, a significant limitation of regression-based methods is that they do not take into account the surrounding context of each keypoint, which can lead to inaccuracies in complex scenarios where multiple keypoints may be closely located or occluded. To address this limitation, heatmap-based approaches [42], [43], [44], [45] have emerged, which localize keypoints by generating probabilistic heatmaps instead of fixed coordinates. These heatmaps provide a richer representation of the keypoint spatial distribution, allowing for better utilization of keypoint surrounding areas. In this paper, our method is also based on heatmaps.

### 2.2 Category-Agnostic Pose Estimation

The task of Category-Agnostic Pose Estimation (CAPE) seeks to develop a pose estimation model capable of detecting the poses of various novel object categories using only a few reference examples. Traditional approaches [9], [10], [11], [12], [13] typically involve training the model on a predefined set of base categories, then adapting it to unseen categories using the given few-shot examples. While effective, these methods often require extensive training data and may struggle to generalize to new categories that differ significantly from those encountered during training. Recent advancements have continued to expand the capabilities of CAPE. GraphCape [46] introduces a graph-based framework that treats keypoints as connected nodes to exploit geometric relationships, improving symmetry breaking and occlusion handling. Building on GraphCape, EdgeCape [47] further predicts adaptive edge weights and incorporates Markovian Structural Bias to better capture global spatial dependencies. SCAPE [48], in contrast, streamlines the design through implicit matching via self-attention, with a global keypoint perceptor and attention refiner enhancing both accuracy and efficiency. In this work, we approach this task from a novel perspective by harnessing the rich knowledge embedded within the off-the-shelf T2I diffusion models. By leveraging the cross-attention mechanism, we can effectively establish correspondences between example images and test images of novel categories, even to the extent of eliminating the need for any training on specific base categories.

### 2.3 Text-to-Image Diffusion Models

Text-to-Image (T2I) diffusion models [18], [19], [20], [21] are powerful generative models that employ diffusion techniques to produce remarkably high-quality images. These models are primarily designed to generate detailed and coherent images based on textual descriptions, allowing for a wide range of creative possibilities. Recently, T2I diffusion models have made remarkable strides across various domains, including image inpainting [49], which aims to fill in missing parts of images; image editing [50], which allows users to modify images in intuitive ways; pixel matching [51], which establishes pixel correspondences between different images; and scene understanding [52], which focuses on analyzing and interpreting visual environments. Additionally, T2I diffusion models have also found applications in various interdisciplinary fields [53], showcasing their versatility and effectiveness. As far as we know, in this paper, we are the first to explore the potential of T2I diffusion models for the challenging category-agnostic pose estimation task, addressing the limitations of traditional methods and opening up a new avenue for research in this area.

### 2.4 Attention Refinement in T2I Diffusion Models

Several recent works have also explored leveraging cross-attention maps from T2I diffusion models with attention refinement. Dataset Diffusion [54] introduces a method to generate synthetic segmentation datasets using text-to-image diffusion. It conditions generation with cross-attention prompts and uses self-attention as a signal filter. iSeg [55] extracts segmentation results from diffusion models through iterative cross-attention refinement. It utilizes entropy-reduced self-attention maps as structural priors to suppress weak or uncertain responses, leading to sharper object boundaries. SLiMe [56] presents a segmentation method using T2I diffusion models, which highlights that raw cross-attention is often noisy, and introduces weighted accumulated self-attention to enhance boundary precision. Different from these segmentation methods, our proposed Foreground-Guided Attention Refinement (FGAR) module is specifically designed for the Category-Agnostic Pose Estimation (CAPE) task, where precise localization of keypoints within the foreground object is critical. Unlike prior methods that focus on boundary-level tuning or require iterative post-processing, our FGAR introduces a foreground-guided and point-concentrated refinement mechanism to directly enhance cross-attention for precise pose estimation.

### 2.5 Prompt Learning

The concept of "prompt" was first introduced in [57] as an instruction appended to the input of large pre-trained models. Prompting has since been widely studied [58], [59], including manually crafted prompts [14], [60], [61], which often underperform when task semantics are difficult to express in words. To overcome this limitation, learnable prompt methods [62], [63], [64] enable models to automatically acquire suitable textual representations, and several works further enhance prompts by injecting external knowledge [65], [66]. Unlike prior methods that learn task prompts for downstream adaptation, we use prompt learning to encode keypoint semantics, enabling T2I diffusion models to tackle the structurally distinct CAPE task.

## 3 Method: Prompt Pose Matching (PPM)

**Task Definition.** Unlike most pose estimation tasks that predict keypoints for a single known (seen) category, Category-Agnostic Pose Estimation (CAPE) requires the model to efficiently generalize to novel (unseen) 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 $D_{\text{base}}$. In $D_{\text{base}}$, we use $I_{\text{base}}$ and $\mathcal{H}_{\text{base}}$ to represent the image and the corresponding keypoint localization label, respectively. Following [9], we use the keypoint labels in heatmap format. For each image $I_{\text{base}}$, its corresponding heatmap label $\mathcal{H}_{\text{base}}$ contains multiple sub-labels, i.e., $\{H^1_{\text{base}}, H^2_{\text{base}}, ..., H^n_{\text{base}}, ..., H^N_{\text{base}}\}$ where each sub-label represents the heatmap of one keypoint, and $N$ denotes the number of keypoints in the image. Previous methods [9], [10], [11], [12] generally train the model on $D_{\text{base}}$ 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 $D_{\text{exm}}$, for few-shot adaptation. In $D_{\text{exm}}$, we use $I_{\text{exm}}$ and $\mathcal{H}_{\text{exm}}$ to denote the example image and its corresponding label. Similarly, $\mathcal{H}_{\text{exm}}$ also contain multiple sub-labels for all keypoints, i,e, $\{H^1_{\text{exm}}, H^2_{\text{exm}}, ..., H^n_{\text{exm}}, ..., H^N_{\text{exm}}\}$. During testing, based on the given few-shot examples $D_{\text{exm}}$, the model is required to detect the corresponding keypoints of the same category for the unlabeled test images $I_{\text{test}}$. 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 share the same number of keypoints.

In this paper, we propose a novel PPM framework that leverages the knowledge in T2I diffusion models for CAPE. Our PPM has good extensibility, which can be applied to various T2I diffusion models. For clarity, we use Stable Diffusion [21] as an example to illustrate our framework.

### 3.1 Category-Agnostic Prompt Pre-training

By delving into the diffusion model's cross-attention layers, we can extract attention maps that are directly conditioned on the input prompt, thereby capturing semantically aligned foreground activations (the details of attention extraction will be described in Sec. 3.2). Motivated by this, we here introduce a Category-Agnostic Prompt Pre-training strategy, aiming to pre-learn a *foreground prompt* that consistently highlights foreground regions across arbitrary objects. By learning from base categories $D_{\text{base}}$, the foreground prompt provides a strong initialization for subsequent few-shot adaptation on novel categories $D_{\text{exm}}$. This design ensures that adaptation learning does not start from random or ambiguous initialization, but instead benefits from a category-agnostic prior that stabilizes optimization in the few-shot setting. However, the key challenge of pre-training is the absence of dense foreground masks—the CAPE task only provides sparse keypoint annotations. To overcome this limitation, we design

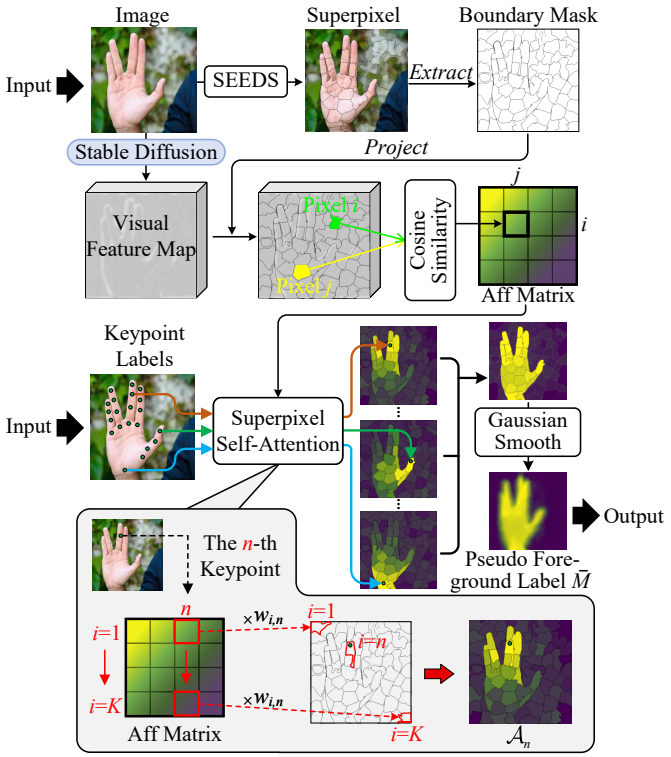



Fig. 3. Architecture of our Foreground-Aware Region Aggregation (FARA) module. FARA converts sparse keypoint annotations into dense pseudo foreground maps. Specifically, FARA constructs Superpixel Self-Attention to highlight regions around each keypoint that are likely to belong to the foreground. By aggregating the superpixel attention maps of all keypoints, FARA produces a complete foreground activation map, which is then used as pseudo supervision for training the foreground prompt.

a Foreground-Aware Region Aggregation (FARA) module that derives foreground masks directly from keypoint labels.

#### 3.1.1 Pre-training Overview

Given a base-category image $I_{\text{base}}$ and its keypoint heatmap labels $H_{\text{base}}$, the FARA module generates a pseudo foreground mask $\bar{M}$ from sparse keypoint annotations. Then, all these masks serve as supervision for training a learnable foreground prompt embedding $P_f$, which interacts with the diffusion model's cross-attention layers. Next, we introduce the FARA module and the foreground prompt learning.

#### 3.1.2 Foreground-Aware Region Aggregation (FARA)

The FARA module generates $\bar{M}$ in the following three steps:

(1) Superpixel Partitioning. As shown in Fig. 3, we first apply the typical SEEDS algorithm [67] to segment the image into $K$ superpixels $\{S_1, S_2, ..., S_K\}$. For each superpixel $S_k$, we average its feature vectors (extracted from diffusion model's intermediate features) to obtain a compact region representation:

$$r_k = \frac{1}{|S_k|} \sum_{x \in S_k} F(x), \tag{1}$$

where $F(x)$ denotes the feature of pixel $x$ from the diffusion model.

(2) Superpixel Self-Attention. Next, we compute cosine similarity between all pairs of superpixel features to construct an affinity matrix:

$$\text{Aff}(i,j) = \frac{r_i^\top r_j}{\|r_i\|_2 \cdot \|r_j\|_2}. \tag{2}$$

where $\|\cdot\|_2$ denotes the L2 norm (or Euclidean norm).

For each keypoint $n$, we use its affinity scores (the corresponding column of the similarity matrix) to generate a superpixel self-attention map $\mathcal{A}_n$, as shown in the bottom panel of Fig. 3.

To further improve the reliability of superpixel self-attention maps, we introduce a distance-based weighting scheme. The intuition is that superpixels closer to a keypoint are more likely to belong to its semantic foreground region, whereas distant superpixels are less relevant and may introduce noise. By assigning higher weights to nearby regions and suppressing far-away activations, we ensure that each superpixel self-attention map better reflects the local semantic structure around the keypoint. Formally, let $c_i$ be the geometric center of superpixel $S_i$, computed as:

$$c_i = \left( \frac{1}{|S_i|} \sum_{(x,y) \in S_i} x, \frac{1}{|S_i|} \sum_{(x,y) \in S_i} y \right), \tag{3}$$

where $u_n$ denote the coordinates of keypoint $n$. For the original attention map, we assign each superpixel a weight:

$$w_{i,n} = \exp\left( - \frac{\|c_i - u_n\|_2^2}{2\delta^2} \right), \tag{4}$$

where $\delta$ controls the spatial decay.

Given the weight $w_{i,n}$, the self-attention map related to keypoint $n$ is formulated as:

$$\mathcal{A}_n(i) = w_{i,n} \cdot \text{Aff}(i,n), \tag{5}$$

where $\mathcal{A}_n(i)$ denotes the value of the $i$-th superpixel of $\mathcal{A}_n$. This is also illustrated at the bottom of Fig. 3.

(3) Attention Aggregation. After obtaining each keypoint-related self-attention map $\{\mathcal{A}_1, \ldots \mathcal{A}_n, \ldots \mathcal{A}_N\}$, we merge them into a complete foreground map as the pseudo label:

$$\bar{M} = \xi\left( \sum_{n=1}^{N} \mathcal{A}_n \right), \tag{6}$$

where $\xi$ denotes applying Gaussian smoothing for spatial harmony.

#### 3.1.3 Foreground Prompt Learning

By leveraging pseudo ground truth $\bar{M}$ obtained from base categories, we optimize the foreground prompt embedding $P_f$ to capture consistent foreground activations. Formally, the pre-training objective is to align the cross-attention map induced by $P_f$, i.e., $CrossAtt(I_{\text{base}}, P_f)$, with the pseudo foreground mask $\bar{M}$:

$$\mathcal{L}_{\text{pre}} = \|CrossAtt(I_{\text{base}}, P_f) - \bar{M}\|^2, \tag{7}$$

where $\|\cdot\|^2$ denotes the $L_2$ distance. This pre-training process ensures that $P_f$ encodes a more generalizable prior, which is crucial for a reliable few-shot adaptation to novel categories.

### 3.2 Few-Shot Prompt Adaptation

After completing pre-training on base categories, we move to few-shot adaptation on novel categories. The foreground

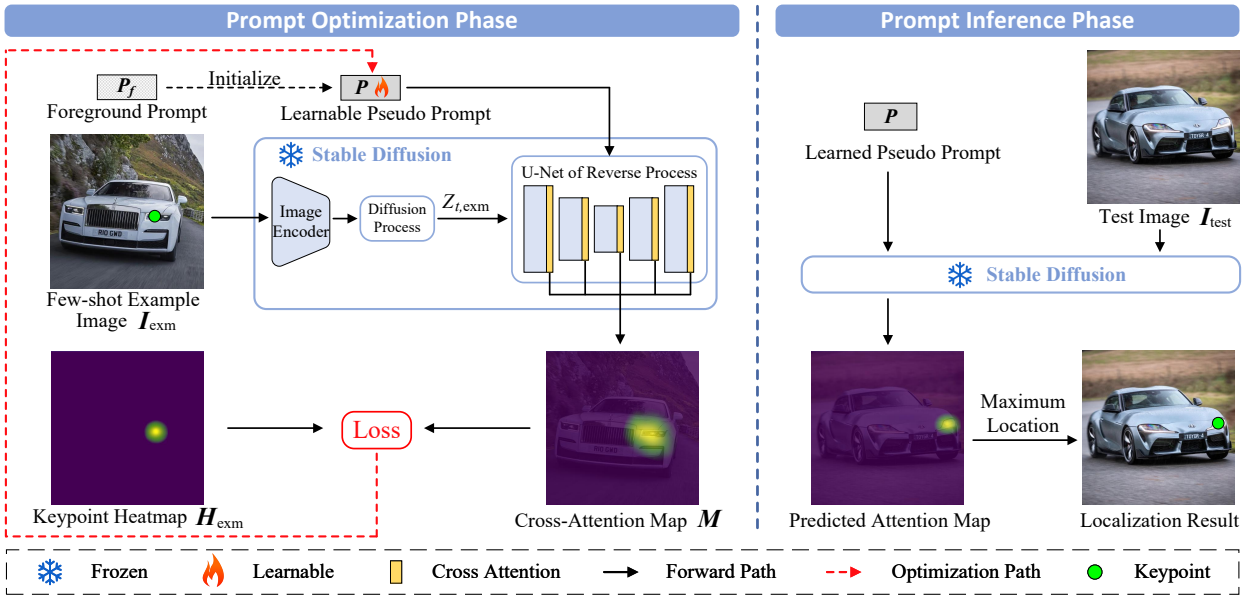



Fig. 4. Few-Shot Prompt Adaptation of our PPM framework. It consists of two major phases. At the prompt optimization phase, given a few-shot example image $I_{\text{exm}}$ and its keypoint heatmap $H_{\text{exm}}$, PPM optimizes a pseudo prompt $P$, which is initialized from the pre-trained foreground prompt $P_f$. At the prompt inference phase, PPM utilizes the learned pseudo prompt $P$ to conduct keypoint detection on the test image $I_{\text{test}}$.

prompt $P_f$ learned in pre-training is adopted as the initialization for adaptation, thereby reducing the difficulty of few-shot learning. Before describing the adaptation procedure, we first provide a brief conceptual clarification. In PPM, each keypoint-specific pseudo prompt functions as a soft textual descriptor whose embedding is optimized so that the resulting cross-attention map aligns with the ground-truth keypoint heatmap in the example image. This perspective makes clear that few-shot prompt adaptation enables PPM to bridge keypoint correspondences from example images to unseen test images.

### 3.2.1 Adaptation Overview

The few-shot adaptation on novel categories consists of two phases. In the first phase (prompt optimization), given an example image $I_{\text{exm}}$ and its keypoint heatmap label $H_{\text{exm}}$, we optimize a pseudo prompt $P$ for each keypoint, using the pre-trained foreground prompt $P_f$ as initialization, as illustrated in Fig. 4. For an example image with $N$ keypoints, $N$ pseudo prompts are optimized accordingly. In the second phase (prompt inference), the optimized pseudo prompts are kept fixed and used to compute $N$ corresponding cross-attention maps for a test image $I_{\text{test}}$. In this way, the learned pseudo prompts bridge keypoint correspondences between example and test images, thereby harnessing Stable Diffusion to address CAPE. Next, we introduce the two phases in detail.

### 3.2.2 Prompt Optimization Phase

In this phase, we aim to learn pseudo prompts to represent the semantics of the keypoints from few-shot examples $D_{\text{exm}}$. Taking the learning for an arbitrary keypoint in the example image as an example, at the beginning, instead of random initialization, a pseudo prompt $P \in \mathbb{R}^c$ is initialized with the parameters of the pre-trained foreground prompt $P_f$, where $c$ is the channel dimension of the embedding space. Given an example image $I_{\text{exm}} \in \mathbb{R}^{H\times W\times 3}$ and its keypoint heatmap $H_{\text{exm}} \in \mathbb{R}^{H\times W}$, we input $P$ and $I_{\text{exm}}$ into Stable Diffusion to obtain cross-attention maps. Specifically, the image encoder and diffusion process convert $I_{\text{exm}}$ into a noisy latent representation $Z_{t,\text{exm}}$, which is then combined with $P$ in the U-Net reverse process. At each U-Net layer, a cross-attention map is produced (yellow regions in Fig. 4), measuring the correlation between $Z_{t,\text{exm}}$ and $P$. Specifically, the cross-attention map at the $l$-th layer in the U-Net can be obtained as:

$$M^l_{\text{CA}} = Sigmoid(\frac{F^l_{Z_{t,\text{exm}}} * F_P^\top}{\sqrt{d_l}}), \tag{8}$$

where $F^l_{Z_{t,\text{exm}}}$ is the visual feature map at $l$-th layer, and $F_P$ is the prompt feature; $F^l_{Z_{t,\text{exm}}}$ and $F_P$ are obtained from $Z_{t,\text{exm}}$ and $P$ respectively; $d_l$ denotes the feature dimension at the $l$-th layer; $Sigmoid$ is the normalization that transforms the values into $[0, 1]$. In Eq. 8, the prompt feature $F_P$ is multiplied with each pixel-level visual feature of the feature map $F^l_{Z_{t,\text{exm}}}$, 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 $M^l_{\text{CA}}$, where $M^l_{\text{CA}} \in \mathbb{R}^{H_l\times W_l}$. 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:

$$M = \frac{1}{L}\sum_{l=1}^{L} M^l_{\text{CA}}, \tag{9}$$

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 $M \in \mathbb{R}^{H\times W}$.

After obtaining the cross-attention map $M$, given the heatmap label $H_{\text{exm}}$ of a keypoint, we calculate the L2 loss between $M$ and $H_{\text{exm}}$, and use the loss to optimize the

pseudo prompt $P$, ensuring that $P$ can drive Stable Diffusion to highlight the specific keypoint region. The optimization loss can be formulated as:

$$\mathcal{L} = \|M - H_{\text{exm}}\|^2, \tag{10}$$

After optimization, we can obtain a learned pseudo prompt $P$ that represents the semantics of a keypoint in example image. By learning from all $N$ keypoint heatmap labels of the example image, we obtain $N$ pseudo prompts. For a novel category with multiple labeled examples (i.e., more than 1-shot), the loss is averaged across examples. Next, we describe how the learned prompts $P$ are used for keypoint localization on the test image $I_{\text{test}}$ (inference stage).

#### 3.2.3 Prompt Inference Phase

After optimization, we use the pseudo prompt $P$ learned from the example image, to find the corresponding keypoint in each test image. As shown in Fig. 4 (right), given a test image $I_{\text{test}}$, we feed $P$ and $I_{\text{test}}$ into Stable Diffusion to compute the cross-attention map. The highlighted region in the output cross-attention map denotes the detected corresponding keypoint region. Finally, we locate the keypoint by finding the position with maximum value in the cross-attention map, as shown in Fig. 4 (right). In this way, based on all learned pseudo prompts, we can estimate the location of all keypoints for the test image of the same category.

### 3.3 Training-Free Attention Enhancement

As mentioned in few-shot adaptation, the learned pseudo prompts are applied to test samples for keypoint localization. During inference, however, the raw cross-attention maps induced by these prompts may still contain background noise or spurious activations, which can degrade localization accuracy. To address this issue without introducing extra training, we further exploit the pre-trained foreground prompt $P_f$ to enhance the quality of attention maps in a training-free manner.

#### 3.3.1 Enhancement Overview

While pseudo prompts enable keypoint-specific localization, the resulting cross-attention maps from Stable Diffusion often remain noisy, especially under cluttered backgrounds. To refine them without additional training, we propose the Foreground-Guided Attention Refinement (FGAR) module, which leverages the pre-trained foreground prompt $P_f$ to generate a foreground attention map $M_f$ for the test image. At test time, the raw cross-attention maps extracted from each U-Net layer are passed through FGAR, yielding the final enhanced attention map for more reliable keypoint localization under the guidance of $M_f$.

#### 3.3.2 Foreground-Guided Attention Refinement (FGAR)

The FGAR module enhances the attention quality in the following two steps:

(1) Foreground-Guided Layer Reweighting. Although cross-attention maps can be obtained from multiple U-Net layers, their quality is not uniform: shallow layers may capture fine-grained local cues but are noisy, whereas deeper layers may encode global semantics but lose spatial precision, as shown in Fig. 5. To adaptively integrate them, we compute a soft IoU score between each layer's attention map $M_{\text{CA}}^l$ and the foreground attention map $M_f$. Note that each attention map $M_{\text{CA}}^l$ is resized to the image resolution for consistency. The soft IoU is formulated as:

$$\text{IoU}_l = \frac{\sum_i \min(M_{\text{CA}}^l(i), M_f(i))}{\sum_i \max(M_{\text{CA}}^l(i), M_f(i))}, \tag{11}$$

where $i$ indexes pixel locations. As shown in Fig. 5, the activated regions in $M_{\text{CA}}^3$ fall outside the target object, resulting in a significantly low IoU score. Since the integration weight depends on the IoU score, the contribution of $M_{\text{CA}}^3$ to the final outcome is marginal.

While IoU directly measures spatial overlap, it alone may be insufficient: an unfocused attention map can still achieve a high IoU but lacks precision for keypoint localization. Take the attention map $M_{\text{CA}}^1$ in Fig. 5 as an example, although it achieves a relatively high IoU score, its attention is scattered, which introduces considerable uncertainty for keypoint localization.

To address this, we also introduce a focus score $f_l$ that measures how concentrated the attention is:

$$f_l = \sigma\left(\frac{\sum_i M_{\text{CA}}^l(i)}{\sum_i \mathbf{1}(M_{\text{CA}}^l(i) > \tau)}\right), \tag{12}$$

where the numerator sums all activations in the attention map, representing the total energy of this layer's cross-attention, the denominator counts the number of pixels above a threshold $\tau$ (we set $\tau$ as the mean activation of the attention map), representing the effective activated area, and $\sigma(\cdot)$ is the sigmoid function to normalize $f_l$ into $[0, 1]$. Intuitively, if an attention map has a large, unfocused activation area, the denominator becomes large and $f_l$ decreases; conversely, if an attention map is small but concentrated, the denominator is small and $f_l$ increases. Therefore, a larger $f_l$ indicates that the layer's attention is more focused, rather than spreading over a broad "blurry" region.

The final layer weight is computed as:

$$w_l = \text{IoU}_l \cdot f_l, \tag{13}$$

which emphasizes both foreground overlap and attention concentration. The enhanced attention map is aggregated across layers as:

$$M' = \frac{\sum_{l=1}^{L} w_l \cdot M_{\text{CA}}^l}{\sum_{l=1}^{L} w_l}. \tag{14}$$

This enables the fused map $M'$ to highlight the semantic area while remaining foreground-consistent, producing a more accurate and robust basis for keypoint localization.

(2) Foreground-Guided Spatial Refinement. After obtaining the fused attention map $M'$ through layer reweighting, we further refine it by enforcing spatial alignment with the foreground prior $M_f$. The key intuition is that semantic keypoints should always lie within the object foreground; therefore, any residual activations outside the foreground region are considered unreliable and should be suppressed. To achieve this, we apply an element-wise multiplication between the fused attention map $M'$ and the foreground map $M_f$:

$$\tilde{M} = M' \odot M_f, \tag{15}$$

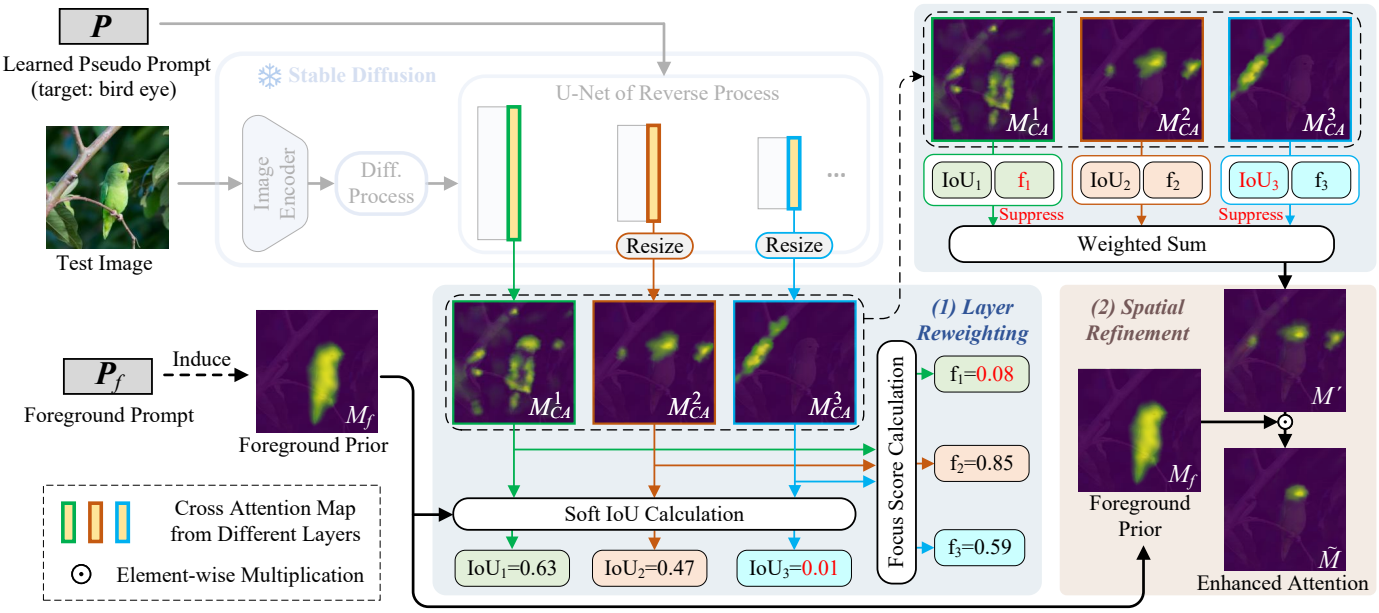



Fig. 5. Architecture of our Foreground-Guided Attention Refinement (FGAR) module. FGAR aggregates cross-attention maps from multiple layers in a training-free manner using the learned foreground prior. It performs two operations: *(1) layer reweighting*, where Soft IoU and Focus Score jointly determine the contribution of each layer, and *(2) spatial refinement*, which suppresses activations outside the foreground region. For clarity, only the first three layers are visualized in the figure.

where $\odot$ denotes element-wise product. This operation preserves activations consistent with the foreground while suppressing spurious responses in the background. The final refined map $\tilde{M}$ thus provides a cleaner and more semantically aligned basis for reliable keypoint localization.

#### 3.3.3 Training-Free Enhancement

In summary, FGAR refines raw cross-attention maps in two steps: (1) layer reweighting guided by foreground consistency and attention sharpness, and (2) spatial refinement enforced by the foreground prior. This entire process can be expressed as a unified training-free refinement operator $\Phi(\cdot)$:

$$\tilde{M} = \Phi(\{M_{\text{CA}}^l\}_{l=1}^{L}, M_f), \tag{16}$$

where FGAR is fully training-free and can be seamlessly integrated into the inference pipeline to produce enhanced attention maps.

### 3.4 Prompt Ensemble Inference

In CAPE, prior methods typically localize one keypoint per inference, requiring $N$ inferences for an image with $N$ keypoints, leading to inefficiency. To address this, we propose a novel Prompt Ensemble Inference (PEI) scheme, enabling our PPM framework to localize multiple keypoints in a single inference pass, significantly improving efficiency. In our PEI, we propose an ensemble prompt that integrates multiple prompts to localize multiple keypoints simultaneously.

At the input space of the T2I diffusion model, after being processed by the text encoder, the text is encoded into a sequence composed of multiple tokens. Each token can guide the diffusion model to independently focus on the corresponding semantics of the image, which is demonstrated in [68], [69]. Based on this, we propose an ensemble prompt $P^* \in \mathbb{R}^{T\times c}$, which is composed of $T$ tokens with each token containing $c$ channels. In this structure, each token is treated as a pseudo prompt corresponding to one keypoint of the image, i.e., $P^* = [P^1, P^2, \ldots, P^T]$. The value of $T$ is fixed and determined by the text encoder of the T2I diffusion model. Given a test image $I_{\text{test}}$ with $N$ keypoints to be predicted, it is important to note that the number of tokens in $P^*$, denoted as $T$, is typically much larger than the number of keypoints in the images, represented as $N$ (i.e., $T > N$). When $P^*$ is fed into the Stable Diffusion model, in each layer, cross-attention is calculated separately for each pseudo prompt (token), resulting in $T$ cross-attention maps per layer. In this way, our PPM can obtain $T$ attention maps in a single forward pass, each corresponding to one of the $T$ prompts.

As mentioned in the Few-Shot Prompt Adaptation stage (Sec. 3.2), all prompts are initialized from the pre-trained foreground prompt $P_f$. Since all prompts share identical initialization, there is no inherent distinction between them. Consequently, for few-shot adaptation on $N$ keypoints of the novel category, we directly optimize the attention maps corresponding to the first $N$ prompts, which is equivalent to optimizing any subset of the whole $T$ prompts due to their identical initialization.

### 3.5 Training and Testing

Our PPM framework can operate in different scenarios, either *with* or *without* data from base categories.

For PPM *with* base categories, we strictly follow the procedures described in Secs. 3.1-3.4 for training and inference. In short, the foreground prompt is first pre-trained on the base categories (Sec. 3.1), followed by few-shot adaptation on the novel categories (Sec. 3.2). During testing, inference is performed using training-free attention enhancement (Sec. 3.3) and prompt ensemble inference (Sec. 3.4).

For PPM *without* base categories, given few-shot labeled examples of a novel category, we perform prompt pre-training (Sec. 3.1) using these examples, which still allows us to train a foreground prompt $P_f$.

TABLE 1
Results on MP-100 in the setting of Cross Super-Category Generalization (1-shot). Parentheses show prompt-learning time.

| Method | Base Cate. | Hum. Body | Hum. Face | Vehicle | Furniture | Avg. Accuracy | Inf. Time (hrs) |
|---|---|---|---|---|---|---|---|
| ProtoNet [10] | ✓ | 37.61 | 57.80 | 28.35 | 42.64 | 41.60 | 4.13 |
| MAML [11] | ✓ | 51.93 | 25.72 | 17.68 | 20.09 | 23.08 | 2.23 |
| Fine-tune [12] | ✓ | 52.11 | 25.53 | 17.46 | 20.76 | 28.97 | 2.05 |
| POMNet [9] | ✓ | 73.82 | 79.63 | 34.92 | 47.27 | 58.91 | 4.82 |
| CapeFormer [13] | ✓ | 83.44 | 80.96 | 45.40 | 52.49 | 65.57 | 4.96 |
| SCAPE [48] | ✓ | 84.24 | 85.98 | 45.61 | 54.13 | 67.49 | 4.17 |
| GraphCape [46] | ✓ | 81.62 | 83.67 | 50.18 | 55.26 | 67.68 | 2.92 |
| EdgeCape [47] | ✓ | 88.51 | **89.62** | 45.55 | **64.80** | 72.12 | 3.17 |
| PPM | ✗ | 88.45 | 85.12 | 54.36 | 61.83 | 72.44 | 0.94 (0.07) |
| | ✓ | **89.17** | 86.32 | **54.79** | 63.16 | **73.29** | **0.91** (0.04) |

TABLE 2
Results on MP-100 in the setting of Cross Sub-Category Generalization (1-shot).

| Method | Base Cate. | Split1 | Split2 | Split3 | Split4 | Split5 | Avg. Accuracy | Inf. Time (hrs) |
|---|---|---|---|---|---|---|---|---|
| ProtoNet [10] | ✓ | 46.05 | 40.84 | 49.13 | 43.34 | 44.54 | 44.78 | 3.67 |
| MAML [11] | ✓ | 68.14 | 54.72 | 64.19 | 63.24 | 57.20 | 61.50 | 1.84 |
| Fine-tune [12] | ✓ | 70.60 | 57.04 | 66.06 | 65.00 | 59.20 | 63.58 | 1.79 |
| POMNet [9] | ✓ | 84.23 | 78.25 | 78.17 | 78.68 | 79.17 | 79.70 | 4.39 |
| CapeFormer [13] | ✓ | 89.45 | 84.88 | 83.59 | 83.53 | 85.09 | 85.31 | 4.42 |
| SCAPE [48] | ✓ | 91.67 | 86.87 | 87.29 | 85.01 | 86.92 | 87.55 | 3.96 |
| GraphCape [46] | ✓ | 91.19 | 87.81 | 85.68 | 85.87 | 85.61 | 87.23 | 2.52 |
| EdgeCape [47] | ✓ | 93.69 | 89.27 | 87.85 | 86.67 | 87.59 | 89.01 | 2.66 |
| PPM | ✗ | 92.95 | 89.93 | 86.66 | 87.86 | 87.89 | 89.06 | 0.85 (0.06) |
| | ✓ | **93.74** | **90.24** | **88.13** | **89.04** | **89.37** | **90.10** | **0.81** (0.02) |

TABLE 3
Results on MP-100 in the setting of Cross Sub-Category Generalization (5-shot).

| Method | Base Cate. | Split1 | Split2 | Split3 | Split4 | Split5 | Avg. Accuracy | Inf. Time (hrs) |
|---|---|---|---|---|---|---|---|---|
| ProtoNet [10] | ✓ | 60.31 | 53.51 | 61.92 | 58.44 | 58.61 | 58.56 | 3.81 |
| MAML [11] | ✓ | 70.03 | 55.98 | 63.21 | 64.79 | 58.47 | 62.50 | 1.96 |
| Fine-tune [12] | ✓ | 71.67 | 57.84 | 66.76 | 66.53 | 60.24 | 64.61 | 1.92 |
| POMNet [9] | ✓ | 84.72 | 79.61 | 78.00 | 80.38 | 80.85 | 80.71 | 4.41 |
| CapeFormer [13] | ✓ | 91.94 | 88.92 | 89.40 | 88.01 | 88.25 | 89.30 | 4.48 |
| SCAPE [48] | ✓ | 93.42 | 89.91 | 90.61 | 89.44 | 89.95 | 90.66 | 4.05 |
| GraphCape [46] | ✓ | 94.24 | 91.32 | 90.15 | 90.37 | 89.73 | 91.16 | 2.56 |
| EdgeCape [47] | ✓ | 95.51 | 91.94 | 91.33 | 90.36 | 91.92 | 92.21 | 2.74 |
| PPM | ✗ | 94.22 | 91.83 | 92.32 | 91.38 | 91.94 | 92.34 | 0.89 (0.12) |
| | ✓ | **95.76** | **93.78** | **92.36** | **92.54** | **92.67** | **93.42** | **0.86** (0.09) |

## 4 EXPERIMENTS

Unless otherwise specified, in this section, PPM denotes our framework that includes all designs by default.

### 4.1 Dataset and Metric.

Most pose estimation datasets are unsuitable for the CAPE task, as they primarily include only a single category. Following prior CAPE research [9], we evaluate our approach on the benchmark MP-100 [9], which is composed of many popular 2D pose estimation datasets, including COCO [24], 300W [70], AFLW [71], OneHand10K [72], DeepFasion2 [73], MacaquePose [74], Vinegar Fly [75], Desert Locust [76], etc. 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 cross-category evaluation.

We also evaluate our framework on the widely-used semantic correspondence dataset, SPair-71k [77], which is a challenging dataset containing diverse variations in viewpoint and scale. Since the training, validation, and testing sets of SPair-71k [77] share the same categories, it shows limited ability in cross-category evaluation. Therefore, we continue to use the training and validation sets of MP-100 [9] to enable cross-category evaluation. In other words, we replace the texting set of MP-100 [9] with a more challenging SPair-71k [77] dataset. We manually remove some categories in the training and validation sets of MP-100, ensuring there is no category overlap for evaluation on SPair-71k.

In this paper, following previous work [9], [10], [11], [12], [13] in category-agnostic pose estimation, we use the standard metric of we report PCK@0.2 to assess the algorithm's accuracy. That is, the PCK (Probability of Correct Keypoint) [78] considers a prediction correct if it falls within a distance of $0.2 \times \max(w, h)$ from the ground-truth location, where $w$ and $h$ denote the width and height of the object bounding box, respectively.

### 4.2 Parameter Setting

During pre-training on base categories, we adopt the Adam optimizer [79] with a learning rate of $1e-3$. During the test-time adaptation on few-shot examples, the learning rate is $5e-4$. In the diffusion process of Stable Diffusion, we use $t = 10$ to obtain $z^t$. The U-Net layer number is $L = 9$ (one

cross-attention map extracted from each convolutional block). In FARA, we set $\delta = 0.1$. In PEI, we follow the structure of Stable Diffusion to set $T = 77$ and $c = 768$.

### 4.3 Memory Usage

All experiments are conducted on two NVIDIA RTX 3090 GPUs (24 GB each) with a per-GPU batch size of 4. During pretraining on base categories, the memory usage is about 16 GB per GPU. This memory usage primarily stems from multi-layer attention supervision and gradient-based optimization. During few-shot adaptation on novel categories, the memory is increased to about 18 GB, due to the parallel optimization of multiple keypoint-specific prompts for each image.

In contrast, inference is memory-efficient, requiring about 5 GB per GPU, even though our model simultaneously processes multiple keypoints within a single pass. This is because our PEI scheme allows U-Net feature sharing across all prompts within a single forward pass. By avoiding repeated feature extraction for each keypoint, our approach eliminates redundant memory allocation.

### 4.4 Experimental Settings

For experiments on dataset MP-100 [9], there are two settings: (1) Cross Super-Category Generalization and (2) Cross Sub-Category Generalization. In Cross Super-Category Generalization, we evaluate the model's ability to generalize across different super-categories, where the model is trained on certain super-categories (i.e., base categories) and tested on others. In Cross Sub-Category Generalization, we assess the model's generalization performance across different sub-categories, where training on some sub-categories (i.e., base categories) and testing on others. 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 [9], [13] conduct experiments under three scenarios: *Cross Super-Category Generalization (1-shot)*, *Cross Sub-Category Generalization (1-shot)*, and *Cross Sub-Category Generalization (5-shot)*, we follow them to conduct experiments under the same 3 scenarios for fair comparisons.

For experiments on dataset SPair-71k [77], we focus on the Cross Sub-Category Generalization setting, as SPair-71k only contains sub-categories. Following previous work [10], [11], [12], we conduct experiments under two few-shot settings: (1) 1-shot and (2) 5-shot, thus establishing two evaluation scenarios: *Cross Sub-Category Generalization (1-shot)* and *Cross Sub-Category Generalization (5-shot)*.

Besides, we also evaluate the *inference time* of different methods. For CAPE methods that involve test-time optimization, the reported inference time includes the cost of test-time learning on the few-shot support samples, in addition to the forward evaluation on all test images. This setting is consistent with the protocol adopted by prior CAPE methods. Specifically, as noted in previous study [9], some methods (such as MAML [11] and Fine-tune [12]) explicitly treat few-shot training as part of the testing phase, since the few-shot samples become available only at inference.

### 4.5 Experimental Results

**Experimental Results on MP-100.** In MP-100 dataset, we follow previous work [9], [13] to evaluate the generalization performance of the model on four super-categories respectively: *human body*, *human face*, *vehicle*, and *furniture*. As shown in Tab. 1, the performance of our proposed PPM obviously surpasses that of previous methods, clearly demonstrating the effectiveness of our framework. Most previous methods tend to achieve suboptimal performance, due to the significant divergence between training categories and the testing one, which hinders the generalization of methods that highly rely on knowledge from 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., even without training on base categories, our PPM outperforms previous methods by 19.26% in average accuracy). When learning from base categories, the performance of our PPM is further boosted, which demonstrates its effect on learning the common foreground knowledge. Tab. 1 also presents the corresponding inference time. Note that the reported inference time includes the few-shot optimization performed on the support samples of novel categories. The same definition applies to all later tables. As shown in Tab. 1, given an image with multiple keypoints to predict, different from previous methods that mainly require inferring many times, our method primarily relies on a single inference. This is because our framework adopts the PEI scheme to simultaneously locate multiple keypoints simultaneously via an ensemble prompt, which significantly lowers the inference time of our method. The values in parentheses denote the prompt optimization time of PPM, which is minimal because optimization is performed only on few-shot examples. We also observe a slight increase in the inference time of PPM when base categories are not used (from 0.91 hrs to 0.94 hrs). This is because, during inference, PPM *without* base categories requires foreground pre-training with test-time few-shot examples.

Following [9], [13], we also evaluate the generalization capability of our framework to unseen sub-categories. The results on 1-shot and 5-shot settings are shown in Tabs. 2 and 3. We follow previous work [9], [13] to conduct the evaluation on five different data splits. We show the results of each split and the mean results averaged over all the five splits. From Tabs. 2 and 3, we can see: (1) our PPM *without* base categories, can outperform previous methods trained on base categories across all few-shot settings. (2) When using PPM *with* base categories, our method achieves further enhanced performance and clearly surpasses existing methods, which demonstrates the effectiveness of foreground prompt pre-training. (3) Regardless of whether base categories are used, our method offers a significant advantage in inference time (notably faster than other approaches), highlighting the effectiveness of (4) The prompt-optimization time cost is small, as learning occurs only on a few examples and does not noticeably affect overall inference efficiency.

**Experimental Results on SPair-71k.** We present our results on SPair-71k in Tabs. 4 and 5. It can be observed that, in both the 1-shot and 5-shot settings, our PPM *without* base categories outperforms previous approaches using base categories. This demonstrates that our approach effectively harnesses the pose knowledge embedded in T2I diffusion models. We report the performance across all 18 (sub)categories of SPair-71k, along with the average

TABLE 4
Results on SPair-71k in the setting of Cross Sub-Category Generalization (1-shot).

| Method | Base Cate. | Aero | Bike | Bird | Boat | Bottle | Bus | Car | Cat | Chair | Cow | Dog | Horse | Motor | Person | Plant | Sheep | Train | TV | Avg. Accuracy | Infer. Time (hrs) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ProtoNet [10] | ✓ | 14.57 | 13.68 | 39.18 | 16.72 | 26.53 | 15.29 | 20.48 | 27.04 | 8.23 | 19.86 | 11.55 | 13.54 | 15.19 | 11.36 | 12.47 | 8.24 | 25.68 | 17.51 | 16.65 | 4.72 |
| MAML [11] | ✓ | 26.17 | 19.95 | 50.36 | 24.41 | 31.72 | 22.86 | 35.74 | 35.63 | 13.75 | 24.06 | 18.39 | 23.67 | 20.33 | 22.12 | 19.48 | 14.52 | 32.87 | 31.20 | 24.73 | 2.68 |
| Fine-tune [12] | ✓ | 30.44 | 27.19 | 54.26 | 28.71 | 35.43 | 26.07 | 38.16 | 42.85 | 15.23 | 29.94 | 27.62 | 24.83 | 23.70 | 32.58 | 28.69 | 21.73 | 35.92 | 37.45 | 31.16 | 2.62 |
| POMNet [9] | ✓ | 48.28 | 35.71 | 66.16 | 35.92 | 40.57 | 30.04 | 43.85 | 56.29 | 20.58 | 35.37 | 38.46 | 27.74 | 31.62 | 35.67 | 36.41 | 29.18 | 40.33 | 44.82 | 38.72 | 5.23 |
| CapeFormer [13] | ✓ | 52.37 | 38.63 | 71.57 | 36.52 | 41.08 | 31.92 | 47.35 | 57.64 | 22.77 | 37.54 | 41.32 | 30.73 | 32.94 | 40.87 | 37.16 | 32.47 | 42.62 | 48.85 | 41.35 | 5.31 |
| SCAPE [48] | ✓ | 49.91 | 38.70 | 69.15 | 37.46 | 40.53 | 28.37 | 48.83 | 58.11 | 23.47 | 37.26 | 40.49 | 29.54 | 33.06 | 39.37 | 36.38 | 28.85 | 43.41 | 47.73 | 40.59 | 5.06 |
| GraphCape [46] | ✓ | 51.67 | 40.62 | 70.26 | 38.63 | 42.85 | 31.75 | 48.16 | 58.89 | 24.27 | 38.57 | 41.05 | 30.81 | 34.86 | 41.22 | 37.94 | 30.46 | 44.61 | 48.40 | 41.95 | 4.27 |
| EdgeCape [47] | ✓ | 51.44 | 41.02 | 71.93 | 38.15 | 44.54 | 32.76 | 49.89 | 59.41 | 26.30 | 39.03 | 42.14 | 32.55 | 34.94 | 42.03 | 38.78 | 30.62 | 45.27 | 49.32 | 42.78 | 4.61 |
| PPM | ✗ | 53.49 | 39.83 | 72.64 | 39.75 | 42.51 | 33.23 | 49.57 | 58.36 | 25.18 | 40.06 | 42.35 | 32.59 | 35.14 | 42.68 | 37.28 | 32.47 | 46.71 | 49.62 | 42.97 | 1.21 (0.06) |
| | ✓ | **56.68** | **43.17** | **73.82** | **41.67** | **46.65** | **35.29** | **51.04** | **61.27** | **29.82** | **41.33** | **44.34** | **35.86** | **37.87** | **44.53** | **40.57** | **37.74** | **47.48** | **52.72** | **45.66** | **1.16** (0.01) |

TABLE 5
Results on SPair-71k in the setting of Cross Sub-Category Generalization (5-shot).

| Method | Base Cate. | Aero | Bike | Bird | Boat | Bottle | Bus | Car | Cat | Chair | Cow | Dog | Horse | Motor | Person | Plant | Sheep | Train | TV | Avg. Accuracy | Inf. Time (hrs) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ProtoNet [10] | ✓ | 22.41 | 18.63 | 48.39 | 23.07 | 32.41 | 24.97 | 31.64 | 35.72 | 13.57 | 27.28 | 18.24 | 20.30 | 22.06 | 19.55 | 18.76 | 11.83 | 31.24 | 23.43 | 24.64 | 4.83 |
| MAML [11] | ✓ | 30.23 | 25.75 | 56.12 | 28.64 | 36.32 | 28.68 | 38.26 | 41.25 | 17.33 | 29.08 | 22.43 | 28.14 | 25.78 | 28.30 | 23.41 | 18.93 | 37.65 | 35.82 | 30.67 | 2.79 |
| Fine-tune [12] | ✓ | 34.56 | 30.93 | 58.48 | 31.27 | 38.66 | 29.62 | 40.53 | 45.81 | 18.14 | 32.39 | 30.26 | 27.73 | 26.87 | 33.97 | 29.73 | 22.02 | 38.40 | 39.67 | 33.84 | 2.74 |
| POMNet [9] | ✓ | 50.38 | 38.17 | 69.74 | 37.65 | 42.07 | 31.97 | 45.78 | 58.12 | 22.01 | 36.37 | 40.15 | 29.47 | 32.03 | 37.18 | 38.27 | 30.28 | 42.64 | 46.82 | 40.51 | 5.28 |
| CapeFormer [13] | ✓ | 53.17 | 40.66 | 72.46 | 38.93 | 42.67 | 32.16 | 49.62 | 60.37 | 23.57 | 39.54 | 42.68 | 32.50 | 34.62 | 41.35 | 39.11 | 33.75 | 44.32 | 50.29 | 42.88 | 5.35 |
| SCAPE [48] | ✓ | 52.47 | 39.48 | 70.98 | 40.59 | 41.60 | 31.29 | 50.34 | 59.63 | 24.77 | 40.16 | 41.81 | 33.05 | 35.39 | 42.37 | 38.58 | 32.65 | 46.74 | 51.54 | 42.97 | 5.28 |
| GraphCape [46] | ✓ | 53.75 | 40.67 | 72.79 | 41.24 | 43.07 | 32.64 | 49.45 | 58.62 | 26.56 | 41.34 | 41.06 | 32.97 | 35.60 | 42.88 | 39.53 | 34.48 | 45.78 | 52.25 | 43.59 | 4.46 |
| EdgeCape [47] | ✓ | 54.21 | 41.73 | 71.85 | 41.47 | 45.19 | 33.64 | 50.06 | 50.31 | 27.46 | 40.78 | 42.52 | 34.62 | 35.29 | 43.68 | 38.90 | 35.75 | 46.16 | 52.18 | 43.66 | 4.72 |
| PPM | ✗ | 55.92 | 41.44 | 72.16 | 42.15 | 45.62 | 33.91 | 51.18 | 60.26 | 28.85 | 41.57 | 43.68 | 35.07 | 36.12 | 44.16 | 39.97 | 36.14 | 47.03 | 51.96 | 44.84 | 1.27 (0.11) |
| | ✓ | **58.92** | **42.65** | **74.73** | **43.92** | **47.31** | **35.27** | **52.12** | **62.35** | **30.52** | **42.73** | **44.78** | **36.83** | **37.54** | **44.91** | **41.83** | **37.92** | **48.74** | **53.76** | **46.49** | **1.23** (0.07) |

TABLE 6
Results comparison for the UDC task. IT denotes inference time (hours), where values in parentheses indicate the prompt-learning cost.

| Method | **SPair-71k** (PCK@0.01) | | **PF-Pascal** (PCK@$\tau$) | | | | | **TSS** (PCK@0.05) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | IT | $\tau$=0.05 | $\tau$=0.10 | $\tau$=0.15 | Avg. | IT | FG3DCar | JODS | Pascal | Avg. | IT |
| VGG+MLS [80] | 27.4 | 0.31 | - | - | - | - | - | - | - | - | - | - |
| DINO+MLS [81], [81] | 31.1 | 0.33 | - | - | - | - | - | - | - | - | - | - |
| ASIC [82] | 36.9 | 0.39 | - | - | - | - | - | - | - | - | - | - |
| CNNGeo [83] | - | - | 41.0 | 69.5 | 80.4 | 63.6 | 0.27 | 90.1 | 76.4 | 56.3 | 74.4 | 0.06 |
| PARN [84] | - | - | - | - | - | - | - | 89.5 | 75.9 | 71.2 | 78.8 | 0.10 |
| GLU-Net [85] | - | - | 42.2 | 69.1 | 83.1 | 64.8 | 0.32 | 93.2 | 73.3 | 71.1 | 79.2 | 0.07 |
| DINOv1 [86] | 33.3 | 0.42 | 41.5 | 62.4 | 72.5 | 58.8 | 0.54 | 82.8 | 73.9 | 53.9 | 72.0 | 0.12 |
| DINOv2 [86] | 55.6 | 0.42 | 56.2 | 77.3 | 83.3 | 72.3 | 0.55 | 93.9 | 69.4 | 57.7 | 77.7 | 0.12 |
| Semantic-GLU-Net [87] | - | - | 48.3 | 72.5 | 85.1 | 68.6 | 0.39 | **95.3** | **82.2** | 78.2 | 85.2 | 0.09 |
| SD-DINO [88] | **64.0** | 0.45 | 73.0 | **86.1** | **91.1** | 83.4 | 0.78 | 94.3 | 73.2 | 60.9 | 79.7 | 0.17 |
| PPM | 63.2 | 0.34 (0.27) | **78.4** | 82.7 | 90.6 | **83.9** | 0.62 (0.36) | 95.0 | 81.8 | **80.7** | **85.8** | 0.15 (0.09) |

performance. Notably, our PPM *with* base categories yields even greater performance gains, achieving the best generalization performance across all 18 categories compared to other methods, thereby underscoring the effectiveness of foreground-aware pre-training. In Tabs. 4 and 5, we also observe that our method requires significantly less inference time than others, while the prompt optimization cost on the few-shot support samples remains minimal. All these results on SPair-71k demonstrate both the efficacy and efficiency of our approach in more challenging scenarios, where images show diverse viewpoints and scales.

**Experimental Results of Methods with Few-Shot Learning.** To ensure a fair comparison with PPM, we additionally equip baseline methods with the *few-shot learning* that aligns with the adaptation mechanism used in PPM. Specifically, (1) the backbone layers of each baseline model are kept frozen; (2) only the small prediction head is fine-tuned using the few-shot support examples of the novel category, mirroring PPM's design that updates only lightweight prompt parameters. (3) fine-tuning is performed for 20 iterations for each few-shot sample, consistent with the number of steps used in PPM's prompt adaptation.

As shown in Tab. 7, the fine-tuned baseline methods exhibit slightly better overall performance after applying few-shot adaptation. Since MAML [11] and Fine-tune [12] inherently employ few-shot learning by design, their performance remains unchanged; all other baselines show slight improvements under this setting. Among these baseline methods, the additional inference-time cost introduced by fine-tuning is negligible, as optimization is performed on only few-shot samples. From Tab. 7, we can see that PPM shows highly competitive accuracy even under this strengthened evaluation protocol, despite not relying on training from base categories. Moreover, when base-category data are incorporated, PPM distinctly achieves the best overall performance. These results highlight the advantage of PPM's prompt-based few-shot learning, which effectively leverages the pose knowledge in T2I diffusion models.

**Extension to Unsupervised Dense Correspondences (UDC).** To demonstrate generality beyond category-agnostic pose estimation, we extend PPM to the unsupervised dense prediction setting of Unsupervised Dense Correspondences (UDC). UDC aims to establish dense pixel-to-pixel correspondences between two images depicting semantically related objects. Following prior work [88], we evaluate on three representative benchmarks: (1) SPair-71k [77], (2) PF-PASCAL [89], and (3) TSS [90]. All of these datasets contain image pairs with ground-truth pixel-to-pixel correspondences. We follow [88] and report PCK@$\tau = 0.01$ on SPair-71k, PCK@$\tau$ with $\tau \in 0.05, 0.10, 0.15$ on PF-PASCAL, and PCK@$\tau = 0.05$ on three TSS subsets (FG3DCar, JODS, PASCAL), as shown in Tab. 6. Among these datasets, TSS [90] provides strict dense correspondences, whereas the other two datasets provide relatively sparse correspondences. Following [88], we conduct experiments on all three datasets to ensure a more comprehensive evaluation. To adapt to the UDC task, PPM learns specific prompts for pixels in the support image, enabling the text-to-image diffusion model to align query pixels to their semantic counterparts. Since foreground masks are provided in UDC, our PPM can still operate effectively on this task. As shown in Tab. 6, our PPM achieves highly competitive PCK results across all three datasets (ranking either first or second), indicating that it generalizes effectively from category-agnostic pose estimation to the broader dense-prediction scenario. Unlike CAPE, where the few-shot support examples can be reused for inferring many test images, the UDC task provides image pairs at test time. Our PPM requires to treat one image in each pair as the support image for prompt learning. As a result,

TABLE 7
Results of all methods with few-shot learning. IT denotes inference time.

| Method | Base Cate. | MP-100 | | | | SPair-71k | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1-shot | | 5-shot | | 1-shot | | 5-shot | |
| | | Avg. Acc. | IT (hrs) | Avg. Acc. | IT (hrs) | Avg. Acc. | IT (hrs) | Avg. Acc. | IT (hrs) |
| ProtoNet [10] | ✓ | 46.67 | 3.68 | 59.93 | 3.83 | 19.72 | 4.72 | 28.60 | 4.84 |
| MAML [11] | ✓ | 61.50 | 1.84 | 62.50 | 1.96 | 24.73 | 2.68 | 30.67 | 2.79 |
| Fine-tune [12] | ✓ | 63.58 | 1.79 | 64.61 | 1.92 | 31.16 | 2.62 | 33.84 | 2.74 |
| POMNet [9] | ✓ | 79.87 | 4.41 | 81.34 | 4.45 | 39.07 | 5.24 | 41.13 | 5.29 |
| CapeFormer [13] | ✓ | 85.70 | 4.43 | 89.54 | 4.51 | 41.84 | 5.32 | 42.95 | 5.36 |
| SCAPE [48] | ✓ | 87.63 | 3.97 | 90.82 | 4.07 | 41.28 | 5.06 | 43.32 | 5.29 |
| GraphCape [46] | ✓ | 87.61 | 2.54 | 91.33 | 2.58 | 42.46 | 4.27 | 43.83 | 4.47 |
| EdgeCape [47] | ✓ | 89.22 | 2.68 | 92.41 | 2.77 | 42.98 | 4.62 | 43.72 | 4.74 |
| PPM | ✗ | 89.06 | 0.85 | 92.34 | 0.89 | 42.97 | 1.21 | 44.84 | 1.27 |
| | ✓ | **90.10** | **0.81** | **93.42** | **0.86** | **45.66** | **1.16** | **46.49** | **1.23** |

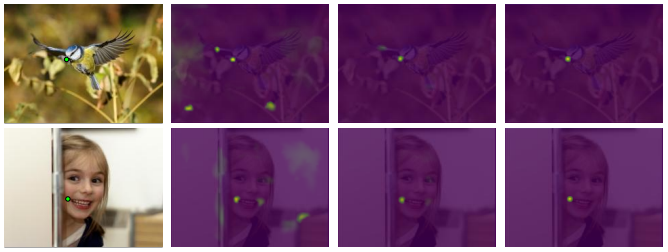



Fig. 6. Visualization of attention maps predicted by our approach.


Fig. 7. Qualitative comparison results of instance swapping in UDC task.


Fig. 8. Representative failure cases of PPM. (a) Failures due to large appearance differences between example and test images. (b) Failures caused by extreme viewpoint changes.

pseudo–prompt optimization must be performed for every test pair, which limits the inference-time advantage that PPM demonstrated in CAPE. Nonetheless, PPM still attains strong performance across diverse UDC benchmarks. In future work, we will investigate how to streamline our PPM to improve the accuracy-efficiency trade-off for UDC.

**Qualitative Results.** In Fig. 9, we qualitatively show the localization results of our method on MP-100 dataset, comparing them to existing CAPE approaches. We can observe that compared to previous methods trained on base categories, our PPM performs much better keypoint localization even when not using data of base categories. Moreover, after using base categories, our PPM gains improved keypoint localization, which demonstrates the effectiveness of our prompt pre-training. We also show the qualitative comparison results on SPair-71k dataset in Fig. 9. We can observe that when the few-shot example largely differs from the test images in terms of viewpoint or scale, our method still performs robustly and generates predictions very close to the ground-truth labels. To demonstrate our FGAR module can effectively improve the quality of attention maps, we also compare the outputted attention maps between our PPM with and without the FGAR module. As shown in Fig. 6, FGAR can drive Stable Diffusion to generate attention maps with cleaner and preciser keypoint regions.

To qualitatively demonstrate the effectiveness of our approach in the UDC task, we follow [88] to leverage instance swapping to visualize the dense correspondence results obtained by different methods. Specifically, instance swapping maps each pixel of the instance object from the target image into the source image according to the obtained dense correspondence results. As shown in Fig. 7, compared with other state-of-the-art methods, our PPM produces smoother results with finer details, highlighting its effectiveness for UDC.

**Failure Cases.** Although PPM performs strongly across benchmarks, it may fail under certain challenging conditions. One limitation arises when the visual appearance of the test image differs substantially from that of the example used for pseudo-prompt learning. Because the pseudo prompts encode the appearance cues present in the example, large discrepancies in shape, material, or design can hinder transferability. As shown in Fig. 8(a), prompts learned from a wooden dining chair do not generalize well to a cushioned armchair; likewise, a decorated bed bears little resemblance to a minimal bed frame, leading to incorrect keypoint localization.

Another failure mode occurs under extreme viewpoint changes. When the example and test images differ drastically in orientation, the model's implicit 3D reasoning becomes insufficient, because the cross-attention mechanism primarily captures 2D appearance correspondences and lacks explicit geometric understanding. As a result, the pseudo prompts learned from the example may activate incorrect regions in the test image. As shown in Fig. 8(b), a bus observed from the front offers little guidance for localizing keypoints on a rear-view bus, and a frontal-view table provides limited support for an overhead-view counterpart.

These cases illustrate the difficulty of guiding keypoint localization from examples to test images when appearance or viewpoint varies widely. Future work may explore stronger structural and 3D-aware priors to further enhance robustness under such conditions.

### 4.6 Ablation Study

**Effect of Pseudo Prompt Learning.** Considering the image keypoints are often difficult to describe textually, we propose to learn pseudo prompts to more appropriately represent the semantics of keypoints. To demonstrate the effectiveness of our prompt learning, we assess the following two variants: (1) We crop the keypoint regions from few-shot examples, input them into an off-the-shelf image caption model [91] to obtain textual descriptions of keypoints, and finally use these descriptions to locate the corresponding keypoints for test images; (2) We crop keypoint regions and send them into the off-the-shelf CLIP's vision encoder [14] to map image regions

into textual embeddings, and finally use these embeddings to find keypoints for test images. Both variants aim to directly describe keypoints using off-the-shelf image-text models, with the only difference being that: the first variant describes keypoints in texts, while the second describes keypoints in textual embeddings. We handle experiments on these two variants using various crop sizes and choose the optimal size for them. As shown in Tab. 8, we can see that our approach, which learns pseudo prompts for keypoints, is much more effective than variants that directly describe keypoints. This is because the semantics of keypoints are often intricate to convey accurately through direct descriptions, while our prompt-learning method allows for a more precise alignment between the learned prompts and the visual keypoints.

TABLE 8
Ablation on pseudo prompt learning.

| Method | MP-100 | | SPair-71k | |
|---|---|---|---|---|
| | 1-shot | 5-shot | 1-shot | 5-shot |
| Keypoint Caption | 26.13 | 27.07 | 15.49 | 17.16 |
| CLIP Encoding | 27.56 | 28.25 | 17.67 | 19.31 |
| Prompt Learning | **90.10** | **93.42** | **45.66** | **46.49** |

**Effect of Foreground Prompt.** In this paper, we pre-train a category-agnostic foreground prompt on base categories, which can activate foreground regions of an image. To verify its effectiveness, we remove all components in PPM that rely on the foreground prompt. Specifically, in the few-shot prompt adaptation stage, we replace the foreground prompt initialization with the widely used random Gaussian initialization. In addition, within FGAR, we remove the parts that utilize the foreground prior. As shown in Tab. 9, removing the foreground prompt leads to a clear performance drop of 3.50% on average, demonstrating that the foreground prompt makes a substantial contribution to achieving the state-of-the-art performance of PPM.

TABLE 9
Ablation on foreground prompt.

| Method | MP-100 | | SPair-71k | |
|---|---|---|---|---|
| | 1-shot | 5-shot | 1-shot | 5-shot |
| PPM (w/o foreground prompt) | 85.94 | 90.65 | 42.16 | 42.93 |
| PPM | **90.10** | **93.42** | **45.66** | **46.49** |

**Effect of FGAR Module.** We conduct an ablation study to demonstrate the effectiveness of our FGAR module. As shown in Tab. 10, compared to directly averaging cross-attention maps, adopting the FGAR module to refine attention maps achieves obviously better performance, leading to 2.27%/2.55%/2.19%/2.15% improvement on four settings respectively, clearly demonstrating the effect of our FGAR.

TABLE 10
Ablation on FGAR module.

| Method | MP-100 | | SPair-71k | |
|---|---|---|---|---|
| | 1-shot | 5-shot | 1-shot | 5-shot |
| PPM (w/o FGAR) | 87.83 | 90.87 | 43.47 | 44.34 |
| PPM | **90.10** | **93.42** | **45.66** | **46.49** |

**Effect of Layer Reweighting and Spatial Refinement in FGAR.** To assess the effectiveness of the components in FGAR, we conduct ablation experiments by removing Layer Reweighting and Spatial Refinement individually. When Layer Reweighting is removed, we simply average the attention maps across layers. The results, reported in Tab. 11, show that removing either Layer Reweighting or Spatial Refinement leads to a clear performance drop of about 0.82%–1.07%, demonstrating the importance of both components. Notably, when both are components removed simultaneously, the performance decreases even further, indicating that the two components are complementary and work together to improve attention quality.

TABLE 11
Ablation on components in FGAR.

| Method | MP-100 | | SPair-71k | |
|---|---|---|---|---|
| | 1-shot | 5-shot | 1-shot | 5-shot |
| PPM | **90.10** | **93.42** | **45.66** | **46.49** |
| PPM (w/o Layer Reweighting) | 89.27 | 92.59 | 44.84 | 45.63 |
| PPM (w/o Spatial Refinement) | 89.11 | 92.46 | 44.50 | 45.42 |
| PPM (w/o both components) | 88.74 | 91.97 | 44.18 | 44.93 |

**FGAR vs. Self-Attention Refinement.** In semantic segmentation, several works based on pre-trained Stable Diffusion models have proposed attention refinement techniques [54], [55], [56]. These methods refine cross-attention with *self-attention* priors, primarily aiming to improve the precision of object boundaries. For a fair comparison, we re-implement these attention refinement methods within our PPM framework by replacing FGAR with each of them in turn, while keeping all other components unchanged.

TABLE 12
Ablation on using different attention refinements.

| Refinement Method | MP-100 | | SPair-71k | |
|---|---|---|---|---|
| | 1-shot | 5-shot | 1-shot | 5-shot |
| No refinement | 88.73 | 91.96 | 42.64 | 43.31 |
| DataDiff [54] | 89.28 | 92.33 | 43.15 | 43.87 |
| iSeg [55] | 89.35 | 92.19 | 43.22 | 43.68 |
| SLiMe [56] | 89.27 | 92.40 | 43.32 | 43.74 |
| FGAR (ours) | **90.10** | **93.42** | **45.66** | **46.49** |
| FGAR+DataDiff | 90.14 | 93.43 | 45.68 | 46.51 |
| FGAR+iSeg | 90.16 | 93.46 | 45.67 | 46.53 |
| FGAR+SLiMe | 90.11 | 93.45 | 45.68 | 46.50 |

As shown in Tab. 12, all three self-attention refinement methods improve over the no-refinement baseline, confirming their effectiveness in enhancing attention quality. While these methods yield noticeable improvements, our FGAR consistently surpasses them on both MP-100 and SPair-71k, highlighting the strength of our design. The performance difference stems from a mismatch in task objectives: self-attention refinements emphasize boundary precision, aligning well with segmentation tasks, but misaligned with pose estimation, where the reliability of the *attention peak* is more critical. In contrast, FGAR produces attention maps that are both foreground-consistent and peak-focused, leading to more reliable and discriminative activations for keypoint localization.

Our FGAR can also incorporate self-attention-based refinement, where the foreground map $M_f$ in FGAR is refined by self-attention to sharpen boundaries. However, each self-attention incorporation yields only marginal improvements (Tab. 12), suggesting that improving the boundary fidelity of the foreground map may be less critical. Given the limited

gains and to avoid methodological redundancy, we do not incorporate self-attention refinement into our FGAR module.

**Effect of PEI Scheme.** In this paper, we propose a PEI scheme to improve our framework's inference efficiency. Here we conduct an ablation study to evaluate the efficiency gains brought by PEI. Compared to our PPM (w/o PEI), which processes one keypoint per inference, the addition of PEI enables our PPM to infer multiple keypoints at once, largely reducing the inference time, as shown in Tab. 13. We can also observe that the introduction of PEI has almost no impact on the model's performance, which indicates that PEI not only enhances inference efficiency but also maintains the model's performance.

TABLE 13
Ablation on PEI scheme.

| Method | MP-100 | | SPair-71k | |
|---|---|---|---|---|
| | Accuracy | Infer. Time (hrs) | Accuracy | Infer. Time (hrs) |
| PPM (w/o PEI) | 90.13 | 4.25 | 45.65 | 5.17 |
| PPM | 90.10 | **0.81** | 45.66 | **1.16** |

**Ablation on Different Focus Score Calculations in FGAR.** In our FGAR module, we introduce a Focus Score to measure the degree of activation concentration for each layer's attention map. To validate the effectiveness of our design, we compare four alternative formulations of this score: (1) Ours (Eq. 12 in the revised paper): ratio of total activation energy to the number of activated pixels (determined by thresholding); (2) Average Activation (AvgAct): mean of all activations in the attention map, without thresholding; (3) Entropy-based Sharpness (Entropy): negative normalized entropy of the attention distribution, where lower entropy indicates more peaked activations; (4) Top-k Ratio (TopK): ratio between the average of top-$k$ activations and the global average activation. Here $k$ is treated as a hyperparameter, and we report the best-performing setting with $k = 15\%$. As shown in Tab. 14, our proposed ratio formulation achieves the highest accuracy across both benchmarks, confirming its advantage over other handcrafted alternatives.

TABLE 14
Ablation on different focus score calculations in FGAR.

| Method | MP-100 | | SPair-71k | |
|---|---|---|---|---|
| | 1-shot | 5-shot | 1-shot | 5-shot |
| Ours | **90.10** | **93.42** | **45.66** | **46.49** |
| AvgAct | 89.16 | 92.79 | 44.32 | 45.83 |
| Entropy | 89.43 | 93.06 | 44.76 | 45.91 |
| TopK | 89.68 | 93.19 | 45.15 | 46.07 |

**Ablation on Different Prompt Initializations.** In PPM, we pre-train a category-agnostic foreground prompt and use it as the initialization for prompts during the few-shot propmpt adaptation stage. To demonstrate the effect of this design, we compare the performance of our PPM under different initialization methods for few-shot adaptation. Specifically, we conduct experiments on the following initialization methods: (1) *Zero Initialization*, where all elements of the prompt embedding are initialized to zero; (2) *Pre-trained Initialization*, where the prompt embedding is initialized with the embedding from a widely-used language encoder, BERT [92]; (3) *Uniform Random Initialization*, where values are randomly drawn from a uniform distribution $[-a, a]$; (4) *Gaussian Random Initialization*, where values are randomly drawn from a normal (Gaussian) distribution. (5) *Foreground Initialization*, which is proposed in this paper. For *Uniform Random Initialization*, we conduct multiple experiments with different values of $a$ and report the best result. As shown in Tab. 15, *Foreground Initialization* consistently outperforms other methods by a clear margin on both datasets, confirming the effectiveness of foreground-aware initialization for keypoint prompt learning.

TABLE 15
Ablation on different prompt initialization methods in PPM.

| Method | MP-100 | | SPair-71k | |
|---|---|---|---|---|
| | 1-shot | 5-shot | 1-shot | 5-shot |
| Zero Initialization | 85.47 | 90.58 | 42.64 | 43.12 |
| Pre-trained Initialization | 86.37 | 90.91 | 42.85 | 43.88 |
| Uniform Rand. Initialization | 86.25 | 91.72 | 43.16 | 43.92 |
| Gaussian Rand. Initialization | 88.76 | 92.08 | 43.51 | 44.27 |
| Foreground Initialization (ours) | **90.10** | **93.42** | **45.66** | **46.49** |

**Extensibility of Our Approach on Other Diffusion Models.** As mentioned in our main paper, the proposed PPM is based on cross-attention to learn pseudo prompts which serve as bridges to build correspondences for CAPE. Since cross-attention is generally used in T2I diffusion models for image generation, our approach can be applied to various T2I diffusion models. We apply our method on the following commonly-used open-source T2I diffusion models: Stable Diffusion [21], Imagen [19], Openjourney [93], and DALLE mini [94]. From Tab. 16, we can see that our PPM consistently outperforms previous methods across all T2I diffusion models, demonstrating its strong extensibility.

TABLE 16
Ablation on using different text-to-image diffusion models.

| Method | MP-100 | | SPair-71k | |
|---|---|---|---|---|
| | 1-shot | 5-shot | 1-shot | 5-shot |
| ProtoNet [10] | 44.78 | 58.56 | 16.65 | 24.64 |
| MAML [11] | 61.50 | 62.50 | 24.73 | 30.67 |
| Fine-tune [12] | 63.58 | 64.61 | 31.16 | 33.84 |
| POMNet [9] | 79.70 | 80.71 | 38.72 | 40.51 |
| CapeFormer [13] | 85.31 | 89.30 | 41.35 | 42.88 |
| SCAPE [48] | 87.55 | 90.66 | 40.59 | 42.97 |
| GraphCape [46] | 87.23 | 91.16 | 41.95 | 43.59 |
| EdgeCape [47] | 89.01 | 92.21 | 42.78 | 43.66 |
| PPM (Stable Diffusion) | 90.10 | 93.42 | 45.66 | 46.49 |
| PPM (Imagen) | 89.94 | 92.67 | 45.41 | 46.83 |
| PPM (Openjourney) | 91.55 | 93.74 | 46.56 | 46.79 |
| PPM (DALLE mini) | 91.82 | 92.97 | 46.72 | 47.15 |

## 5 Conclusion

In this paper, we propose a novel PPM framework to tackle the CAPE task by harnessing the rich structural knowledge embedded in T2I diffusion models. Specifically, PPM introduces three novel components: a Foreground-Aware Region Aggregation (FARA) module that derives pseudo foreground regions from sparse keypoints to enable category-agnostic prompt pretraining, a Foreground-Guided Attention Refinement (FGAR) module that leverages a learned foreground prior to suppress background noise and sharpen keypoint activations, and a Prompt Ensemble Inference (PEI) scheme that integrates multiple prompts for efficient multi-keypoint prediction.

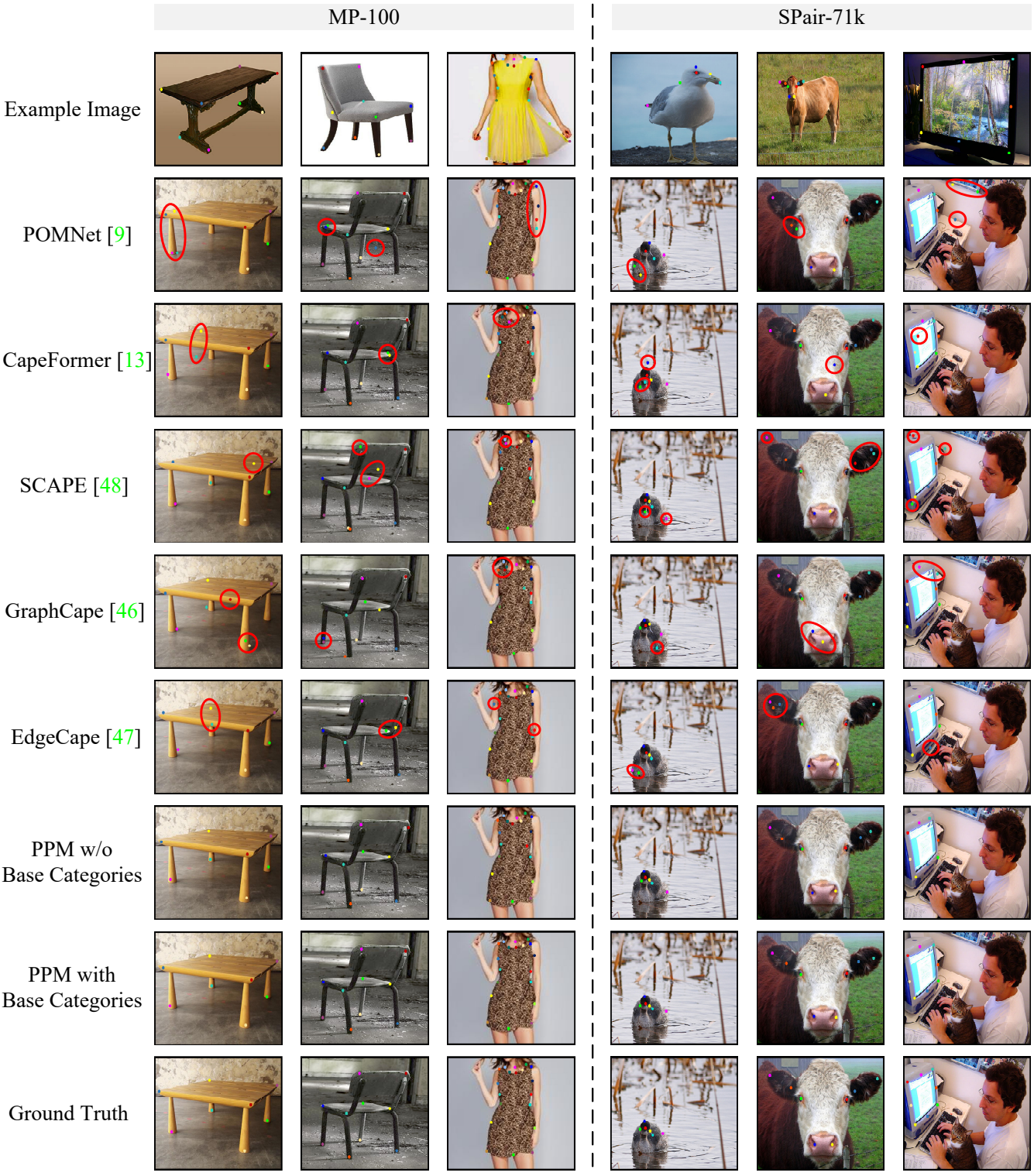



Fig. 9. Qualitative comparison results in CAPE task on both MP-100 and SPair-71k datasets. The keypoints in red circles are failure cases.

## REFERENCES


[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[2] O. G. Guleryuz and C. Kaeser-Chen, "Fast lifting for 3d hand pose estimation in ar/vr applications," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 106–110.

[3] S. Gilroy, M. Glavin, E. Jones, and D. Mullins, "Pedestrian occlusion level classification using keypoint detection and 2d body surface area estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3833–3839.

[4] S. Iftikhar, Z. Zhang, M. Asim, A. Muthanna, A. Koucheryavy, and A. A. Abd El-Latif, "Deep learning-based pedestrian detection in autonomous vehicles: Substantial issues and challenges," *Electronics*, vol. 11, no. 21, p. 3551, 2022.

[5] T. Probst, A. Fossati, and L. Van Gool, "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*. Springer, 2016, pp. 285–301.

[6] W. Xu, P.-c. Su, and S. C. Sen-ching, "Human pose estimation using two rgb-d sensors," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 1279–1283.

[7] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[8] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3d human pose estimation with spatial and temporal transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 656–11 665.

[9] L. Xu, S. Jin, W. Zeng, W. Liu, C. Qian, W. Ouyang, P. Luo, and X. Wang, "Pose for everything: Towards category-agnostic pose estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 398–416.

[10] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[11] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[12] A. Nakamura and T. Harada, "Revisiting fine-tuning for few-shot learning," *arXiv preprint arXiv:1910.00216*, 2019.

[13] M. Shi, Z. Huang, X. Ma, X. Hu, and Z. Cao, "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*, 2023, pp. 7308–7317.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[15] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proceedings of the 2017 ACM on Multimedia Conference*, 2017, pp. 154–162.

[16] Z. Wang, Z. Gao, Y. Yang, G. Wang, C. Jiao, and H. T. Shen, "Geometric matching for cross-modal retrieval," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 3, pp. 5509–5521, 2025.

[17] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based LSTM and semantic consistency," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.

[18] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162. PMLR, 17–23 Jul 2022, pp. 16 784–16 804.

[19] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022.

[20] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[22] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[23] D. Peng, Z. Zhang, P. Hu, Q. Ke, D. Yau, and J. Liu, "Harnessing text-to-image diffusion models for category-agnostic pose estimation," in *European Conference on Computer Vision. Springer*, 2024.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[25] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660.

[26] F. Shen, C. Shen, W. Liu, and H. T. Shen, "Supervised discrete hashing," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 37–45.

[27] F. Shen, Y. Xu, L. Liu, Y. Yang, Z. Huang, and H. T. Shen, "Unsupervised deep hashing with similarity-adaptive and discrete optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 3034–3044, 2018.

[28] A. Bulat and G. Tzimiropoulos, "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*. Springer, 2016, pp. 717–732.

[29] ——, "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*, 2017, pp. 1021–1030.

[30] Z. Liu, Z. Chen, J. Bai, S. Li, and S. Lian, "Facial pose estimation by deep learning from label distributions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[31] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3476–3483.

[32] C. Zimmermann and T. Brox, "Learning to estimate 3d hand pose from single rgb images," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4903–4911.

[33] L. Ge, Y. Cai, J. Weng, and J. Yuan, "Hand pointnet: 3d hand pose estimation using point sets," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8417–8426.

[34] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan, "3d hand shape and pose estimation from a single rgb image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 833–10 842.

[35] N. D. Reddy, M. Vo, and S. G. Narasimhan, "Occlusion-net: 2d/3d occluded keypoint localization using graph networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7326–7335.

[36] C. Zhao, C. Fu, J. M. Dolan, and J. Wang, "L-shape fitting-based vehicle pose estimation and tracking using 3d-lidar," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 4, pp. 787–798, 2021.

[37] H. C. Sánchez, A. H. Martínez, R. I. Gonzalo, N. H. Parra, I. P. Alonso, and D. Fernandez-Llorca, "Simple baseline for vehicle pose estimation: Experimental validation," *IEEE Access*, vol. 8, pp. 132 539–132 550, 2020.

[38] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5933–5942.

[39] J. Wang, X. Long, Y. Gao, E. Ding, and S. Wen, "Graph-pcnn: Two stage human pose estimation with graph pose refinement," in *European Conference on Computer Vision*. Springer, 2020, pp. 492–508.

[40] D. Zhang, G. Guo, D. Huang, and J. Han, "Poseflow: A deep motion representation for understanding human behaviors in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6762–6770.

[41] J. Gong, Z. Fan, Q. Ke, H. Rahmani, and J. Liu, "Meta agent teaming active learning for pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 079–11 089.

[42] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh, "Pose machines: Articulated pose estimation via inference machines," in *European Conference on Computer Vision*. Springer, 2014, pp. 33–47.
[43] W. Tang and Y. Wu, "Does learning specific features for related parts help human pose estimation?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1107–1116.
[44] J. Zhang, Y. Cai, S. Yan, J. Feng *et al.*, "Direct multi-view multi-person 3d pose estimation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 153–13 164, 2021.
[45] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.
[46] O. Hirschorn and S. Avidan, "A graph-based approach for category-agnostic pose estimation," in *European Conference on Computer Vision*. Springer, 2024, pp. 469–485.
[47] ——, "Edge weight prediction for category-agnostic pose estimation," *arXiv preprint arXiv:2411.16665*, 2024.
[48] Y. Liang, Z. Ye, W. Liu, and H. Lu, "Scape: A simple and strong category-agnostic pose estimator," in *European Conference on Computer Vision*. Springer, 2024, pp. 478–494.
[49] S. Xie, Z. Zhang, Z. Lin, T. Hinz, and K. Zhang, "Smartbrush: Text and shape guided object inpainting with diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 428–22 437.
[50] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6007–6017.
[51] L. Tang, M. Jia, Q. Wang, C. P. Phoo, and B. Hariharan, "Emergent correspondence from image diffusion," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 1363–1389.
[52] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9298–9309.
[53] L. X. Nguyen, P. S. Aung, H. Q. Le, S.-B. Park, and C. S. Hong, "A new chapter for medical image generation: The stable diffusion method," in *2023 International Conference on Information Networking (ICOIN)*. IEEE, 2023, pp. 483–486.
[54] Q. Nguyen, T. Vu, A. Tran, and K. Nguyen, "Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 76 872–76 892, 2023.
[55] L. Sun, J. Cao, J. Xie, F. S. Khan, and Y. Pang, "iseg: An iterative refinement-based framework for training-free segmentation," *arXiv preprint arXiv:2409.03209*, 2024.
[56] A. Khani, S. A. Taghanaki, A. Sanghi, A. M. Amiri, and G. Hamarneh, "Slime: Segment like me," *arXiv preprint arXiv:2309.03179*, 2023.
[57] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel, "Language models as knowledge bases?" *arXiv preprint arXiv:1909.01066*, 2019.
[58] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, "How can we know what language models know?" *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 423–438, 2020.
[59] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh, "Autoprompt: Eliciting knowledge from language models with automatically generated prompts," *arXiv preprint arXiv:2010.15980*, 2020.
[60] N. Poerner, U. Waltinger, and H. Schütze, "Bert is not a knowledge base (yet): Factual knowledge vs," *Name-Based Reasoning in Unsupervised QA. CoRR, abs*, 1911.
[61] F. Petroni, P. Lewis, A. Piktus, T. Rocktäschel, Y. Wu, A. H. Miller, and S. Riedel, "How context affects language models' factual predictions," *arXiv preprint arXiv:2005.04611*, 2020.
[62] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.
[63] ——, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 816–16 825.
[64] Y. Zang, W. Li, K. Zhou, C. Huang, and C. C. Loy, "Unified vision and language prompt learning," *arXiv preprint arXiv:2210.07225*, 2022.
[65] B. Kan, T. Wang, W. Lu, X. Zhen, W. Guan, and F. Zheng, "Knowledge-aware prompt tuning for generalizable vision-language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 670–15 680.
[66] H. Yao, R. Zhang, and C. Xu, "Visual-language prompt tuning with knowledge-guided context optimization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 6757–6767.
[67] M. Van den Bergh, X. Boix, G. Roig, B. De Capitani, and L. Van Gool, "Seeds: Superpixels extracted via energy-driven sampling," in *European conference on computer vision*. Springer, 2012, pp. 13–26.
[68] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022.
[69] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or, "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–10, 2023.
[70] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image and vision computing*, vol. 47, pp. 3–18, 2016.
[71] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, "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)*. IEEE, 2011, pp. 2144–2151.
[72] Y. Wang, C. Peng, and Y. Liu, "Mask-pose cascaded cnn for 2d hand pose estimation from single color image," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 11, pp. 3258–3268, 2018.
[73] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "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*, 2019, pp. 5337–5345.
[74] R. Labuguen, J. Matsumoto, S. B. Negrete, H. Nishimaru, H. Nishijo, M. Takada, Y. Go, K.-i. Inoue, and T. Shibata, "Macaquepose: a novel "in the wild" macaque monkey pose dataset for markerless motion capture," *Frontiers in behavioral neuroscience*, vol. 14, p. 581154, 2021.
[75] T. D. Pereira, D. E. Aldarondo, L. Willmore, M. Kislin, S. S.-H. Wang, M. Murthy, and J. W. Shaevitz, "Fast animal pose estimation using deep neural networks," *Nature methods*, vol. 16, no. 1, pp. 117–125, 2019.
[76] J. M. Graving, D. Chae, H. Naik, L. Li, B. Koger, B. R. Costelloe, and I. D. Couzin, "Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning," *Elife*, vol. 8, p. e47994, 2019.
[77] J. Min, J. Lee, J. Ponce, and M. Cho, "Spair-71k: A large-scale benchmark for semantic correspondence. arxiv prepreint," *arXiv preprint arXiv:1908.10543*, vol. 6, pp. 12–14, 2019.
[78] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2878–2890, 2012.
[79] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[80] K. Aberman, J. Liao, M. Shi, D. Lischinski, B. Chen, and D. Cohen-Or, "Neural best-buddies: Sparse cross-domain correspondence," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–14, 2018.
[81] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.
[82] K. Gupta, V. Jampani, C. Esteves, A. Shrivastava, A. Makadia, N. Snavely, and A. Kar, "Asic: Aligning sparse in-the-wild image collections," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4134–4145.
[83] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6148–6157.
[84] S. Jeon, S. Kim, D. Min, and K. Sohn, "Parn: Pyramidal affine regression networks for dense semantic correspondence," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 351–366.
[85] P. Truong, M. Danelljan, and R. Timofte, "Glu-net: Global-local universal network for dense flow and correspondences," in *Proceedings*

*of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6258–6268.
[86] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel, "Deep vit features as dense visual descriptors," *arXiv preprint arXiv:2112.05814*, vol. 2, no. 3, p. 4, 2021.
[87] P. Truong, M. Danelljan, F. Yu, and L. Van Gool, "Warp consistency for unsupervised learning of dense correspondences," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 346–10 356.
[88] J. Zhang, C. Herrmann, J. Hur, L. Polania Cabrera, V. Jampani, D. Sun, and M.-H. Yang, "A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence," *Advances in Neural Information Processing Systems*, vol. 36, pp. 45 533–45 547, 2023.
[89] B. Ham, M. Cho, C. Schmid, and J. Ponce, "Proposal flow: Semantic correspondences from object proposals," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 7, pp. 1711–1725, 2017.
[90] T. Taniai, S. N. Sinha, and Y. Sato, "Joint recovery of dense correspondence and cosegmentation in two images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4246–4255.
[91] J. C. Hu, R. Cavicchioli, and A. Capotondi, "Expansionnet v2: Block static expansion in fast end to end training for image captioning," *arXiv preprint arXiv:2208.06551*, 2022.
[92] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[93] "Openjourney, access date: 8th, nov." https://huggingface.co/prompthero/openjourney, 2022.
[94] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.

**Duo Peng** is a Research Fellow at Institute for Digital Molecular Analytics and Science (IDMxS), Nanyang Technological University. He received the B.E. degree and the M.S. degree from Sichuan University, and the PhD degree from Singapore University of Technology and Design. His main research interests include computer vision, transfer learning and generative AI.

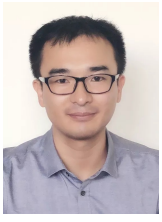
**Zhengbo Zhang** received his Bachelor's and Master's degrees in engineering from Wuhan University, and is currently pursuing a Ph.D. degree at Singapore University of Technology and Design. His current research interest is in computer vision.

**Ping Hu** (Member, IEEE) received the BEng degree from Sichuan University, China, the MEng degree from the University of Chinese Academy of Sciences, China, and the PhD degree from Boston University, USA. He is a professor with the School of Computer Science, the University of Electronic Science and Technology of China. His research interests include machine learning and computer vision. He serves as a senior program committee member for IJCAI 2023 and an area chair for CVPR 2024.

**Qiuhong Ke** is an ARC DECRA Fellow and a Senior Lecturer at the Faculty of Information Technology, Monash University. Previously, she was a Postdoctoral Researcher at Max Planck Institute for Informatics from 2018 to 2019 and a Lecturer at The University of Melbourne from 2020 to 2022. Her research interests include machine learning and computer vision.

**De Wen Soh** received the B.S. degree in Mathematics from Stanford University. He received his Ph.D. degree in Electrical Engineering from Yale University under the supervision of Sekhar Tatikonda, where he worked on high-dimensional graphical model learning. In 2016, he joined the Institute of High Performance Computing, where he worked on machine learning research in relation to social and psychological sciences, alongside various industry projects associated with the consumer and transport industries. His research areas include graphical model estimation, graph signal processing, network analytics, federated learning, transport modelling, high-dimensional statistical theory and artificial intelligence.

**Mohammed Bennamoun** (Senior Member, IEEE) is Winthrop Professor in the Department of Computer Science and Software Engineering at UWA and is a researcher in computer vision, machine/deep learning, robotics, and signal/speech processing. He has published 4 books (available on Amazon), 1 edited book, 1 Encyclopedia article (by invitation), 14 book chapters, 200+ journal papers, 250+ conference publications, 16 invited & keynote publications. His h-index is 74 and his number of citations is 28,000+ (Google Scholar). He was awarded 80+ competitive research grants (approx. 35+ million in funding) from the Australian Research Council, and numerous other Government, UWA and industry Research Grants. He has delivered conference tutorials at major conferences, including IEEE Computer Vision and Pattern Recognition (CVPR 2016), Interspeech 2014, IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) and European Conference on Computer Vision (ECCV). He served for two terms (3 years each term) on the Australian Research Council (ARC) College of Experts, and the ARC ERA 2018. He is currently Senior Area Editor of the IEEE Signal Processing Letters and Associate Editor of the IEEE Transactions on Image Processing and IEEE Transactions on Artificial Intelligence.



**Jun Liu** (Senior Member, IEEE) is a Professor and Chair in Digital Health at School of Computing and Communications in Lancaster University. He got the PhD degree from Nanyang Technological University in 2019. He was with Singapore University of Technology and Design from 2019 to 2024. He is an Associate Editor of IEEE Transactions on Image Processing, IEEE Transactions on Industrial Informatics, IEEE Transactions on Biometrics, Behavior and Identity Science, ACM Computing Surveys, and Pattern Recognition. He has served as an Area Chair of CVPR, ECCV, ICML, NeurIPS, ICLR and MM. His research interests include digital health, computer vision and machine learning.