

# Pose for Everything: Towards Category-Agnostic Pose Estimation

Lumin Xu<sup>1,2\*</sup>, Sheng Jin<sup>3,4\*</sup>, Wang Zeng<sup>1,2</sup>, Wentao Liu<sup>4,5</sup>, Chen Qian<sup>4</sup>,  
Wanli Ouyang<sup>5,6</sup>, Ping Luo<sup>3</sup>, and Xiaogang Wang<sup>1</sup>

<sup>1</sup> The Chinese University of Hong Kong    <sup>2</sup> SenseTime Research  
<sup>3</sup> The University of Hong Kong    <sup>4</sup> SenseTime Research and Tetras.AI  
<sup>5</sup> Shanghai AI Laboratory    <sup>6</sup> The University of Sydney  
{luminxu, zengwang}@link.cuhk.edu.hk    js20@connect.hku.hk  
{liuwentao, qianchen}@sensetime.com    wanli.ouyang@sydney.edu.au  
pluo@cs.hku.hk    xgwang@ee.cuhk.edu.hk

**Abstract.** Existing works on 2D pose estimation mainly focus on a certain category, *e.g.* human, animal, and vehicle. However, there are lots of application scenarios that require detecting the poses/keypoints of the unseen class of objects. In this paper, we introduce the task of Category-Agnostic Pose Estimation (CAPE), which aims to create a pose estimation model capable of detecting the pose of any class of object given only a few samples with keypoint definition. To achieve this goal, we formulate the pose estimation problem as a keypoint matching problem and design a novel CAPE framework, termed POse Matching Network (POMNet). A transformer-based Keypoint Interaction Module (KIM) is proposed to capture both the interactions among different keypoints and the relationship between the support and query images. We also introduce Multi-category Pose (MP-100) dataset, which is a 2D pose dataset of 100 object categories containing over 20K instances and is well-designed for developing CAPE algorithms. Experiments show that our method outperforms other baseline approaches by a large margin. Codes and data are available at <https://github.com/luminxu/Pose-for-Everything>.

**Keywords:** 2D pose estimation, class-agnostic, few-shot, MP-100 dataset

## 1 Introduction

2D pose estimation (also referred to as keypoint localization) aims to predict the locations of the pre-defined semantic parts of an instance. It has received great attention in the computer vision community in recent years because of its broad application scenarios in both academia and industry. For example, human pose estimation [2] has been widely used in virtual reality (VR) and augmented reality (AR); animal pose estimation [67] is of great significance in zoology and wildlife conservation; vehicle pose estimation [44] is critical for autonomous driving.

---

\* indicates equal contribution.

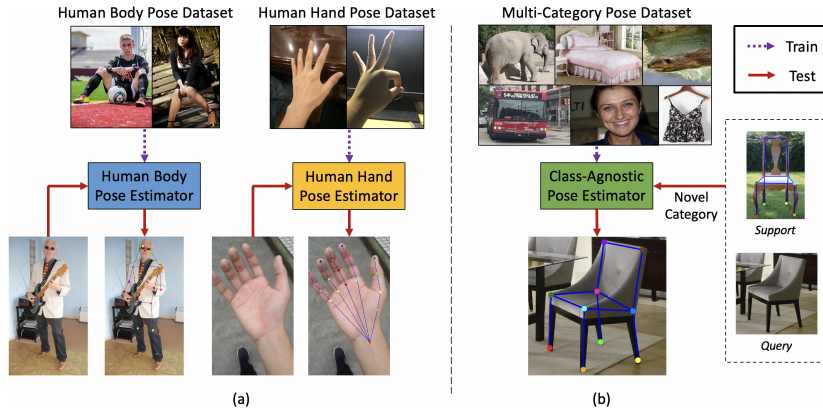


Fig. 1: Category-Specific Pose Estimation vs Class-Agnostic Pose Estimation (CAPE). (a) Traditional pose estimation task is category-specific. Pose estimators are trained on the dataset containing objects of a single category, and can only predict the poses of that category. (b) CAPE task requires the pose estimator to detect poses of arbitrary category given the keypoint definition. After training on the pose dataset containing multi-category objects, the pose estimators can generalize to novel categories given one or a few support images.

The real-world applications from different fields often involve detecting the poses of a variety of novel objects of interest. For example, biologists may study the plant growth by analyzing the poses of plants. However, traditional pose estimators are category-specific and can only be applied to the category that they are trained on. In order to detect poses of novel objects, users have to collect a huge amount of labeled data and design category-specific pose estimation models, which is time-consuming and laborious. To make matters worse, data collection for rare objects (*e.g.* endangered animals) and semantic keypoint annotation for cases that need domain knowledge (*e.g.* medical images) are extremely challenging. Therefore, there is increasing demand for developing pose estimation approaches that can generalize across different categories.

In this paper, we introduce an important yet challenging task, termed Category-Agnostic Pose Estimation (CAPE). As shown in Fig. 1, unlike traditional pose estimation methods that can only predict the poses of a specific category, CAPE aims at using a single model for detecting poses of any category. Given a support image of a novel category and the corresponding keypoint definition, the class-agnostic pose estimator predicts the pose of the same category in the query image. In this way, the pose of any object of interest can be generated according to the arbitrary keypoint definition. The huge cost of data collection, model training and parameter tuning for each novel class is greatly reduced.

There are several challenges preventing the computer vision community from designing systems capable of predicting the poses of a large number of object categories. First, most pose estimation approaches [50] treat it as a supervised

regression task, requiring thousands of labeled images to learn to map an input image to keypoint locations. Second, different objects may have different keypoint definition and unknown number of keypoints. It is non-trivial to learn the unique output representations and utilize the structural information. Third, there are few to none large-scale pose estimation datasets with many visual categories for the development of a general pose estimation method. Previous datasets mostly consist of only one category (*e.g.* human body).

In this paper, we take the first step towards CAPE and propose a novel framework, termed POse Matching Network (POMNet). POMNet formulates the 2D pose estimation task as a matching problem. The keypoint features are extracted from the support images based on the reference keypoint definition, and the image features are extracted from the query image. Matching Head (MH) is designed, which integrates the support keypoint features and the query image features, to estimate the keypoint positions with the maximal possibility. In this way, the model is agnostic to the object category and can be used for any number of keypoints. A transformer-based Keypoint Interaction Module (KIM) is also proposed to capture both the connections among different keypoints and the relationship between the support and query images. The features of different keypoints mutually interact with each other to learn their inherent structure for the given object category. The keypoint features are further aligned with the query image features for better matching. Experimental results show that our model significantly outperforms the other baseline models by a large margin.

In order to train and evaluate the class-agnostic pose estimators, we collect a large-scale pose dataset called Multi-category Pose (MP-100) dataset. The dataset contains over 20K instances, covering 100 sub-categories (*e.g.* vinegar fly body, sofa, suv, and skirt) and 8 super-categories (*e.g.* animal face, furniture, vehicle, and clothes). To our best knowledge, it is the first benchmark that contains the pose annotation of multiple visual (super-)categories.

The main contributions of our work are three-folds.

- We introduce an important yet challenging task termed Category-Agnostic Pose Estimation (CAPE). CAPE requires the model to predict the poses of any objects given a few support images with keypoint definition.
- We propose the novel CAPE framework, namely POse Matching Network (POMNet), and formulate the keypoint detection task as a matching problem. Keypoint Interaction Module (KIM) is proposed to capture both the keypoint-level relationship and the support-query relationship.
- We build the first large-scale multi-(super-)category dataset for the task of CAPE, termed Multi-category Pose (MP-100), to boost the related research.

## 2 Related Works

### 2.1 2D Pose Estimation

There are two types of keypoints in computer vision community. Semantic points are points with clear semantic meanings (*e.g.* the left eye), while interest points

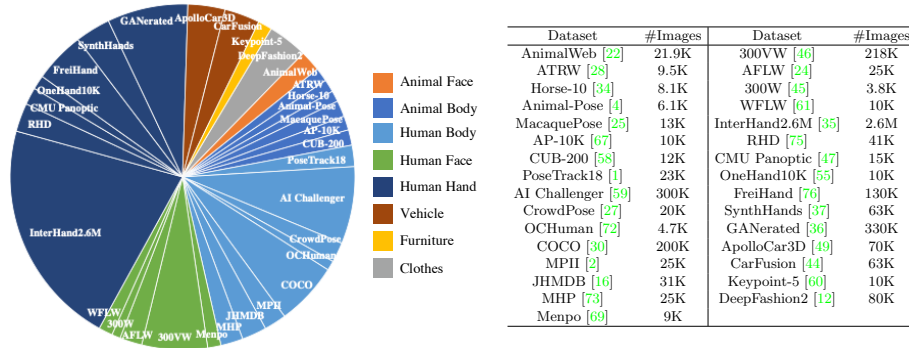


Fig. 2: Categories and image numbers for popular 2D pose estimation datasets.

are low-level points (*e.g.* corner points). 2D pose estimation focuses on predicting the semantic points of objects, *e.g.* human body parts [10,21,30], facial landmarks [3], hand keypoints [75], and animal poses [4]. However, current pose estimation methods and datasets [12,30,67] only focus on keypoints of a single super-category and can not support cross-category/unseen pose estimation.

**2D Pose Estimation Method.** Existing methods can be classified into two categories: regression-based methods [26,40,51,52] and heatmap-based methods [6,7,8,18,19,20,27,39,50,57,62]. Regression-based approaches directly map the image to keypoint coordinates. Such methods are flexible and efficient for real-time applications. However, they are vulnerable to occlusion and motion blur, resulting in inferior performance. Heatmap-based approaches use likelihood heatmaps to encode the keypoint location. Because of excellent localization precision, heatmap-based methods are dominant in the field of 2D pose estimation. Recent works on pose estimation mostly focus on designing powerful convolutional neural networks [6,8,39,50,57,62,63] or transformer-based architectures [29,33,65,68,70]. However, they only focus on detecting the keypoints of object categories that appear during training. In comparison, our model is capable of detecting the keypoints of arbitrary objects of unseen classes.

**2D Pose Estimation Benchmark.** Existing 2D pose estimation datasets only focus on a single super-category. As shown in Fig. 2, most attentions have been focused on human-related categories (*e.g.* human body [1,2,16,27,30,59,72,73], human face [24,45,46,61,69], and human hand [35,36,37,47,55,75,76]), and there are numerous large-scale datasets for these classes. For other long-tailed categories, the datasets are relatively limited in terms of both the dataset sizes and diversity. Nevertheless, analyzing these long-tailed object categories is of great significance in both academia and industry. For example, vehicle pose estimation [44,49] is important for autonomous driving. Animal pose estimation [4,25,28,34,58,67] is of great significance in zoology and wildlife conservation. Indoor furniture pose estimation [60] is important for developing household robots. In this paper, we build the first large-scale benchmark (MP-100 dataset) that contains the pose annotations of a wide range of visual super-categories.

## 2.2 Category-agnostic Estimation

Category-agnostic estimation has been applied to many computer vision tasks, including detection [17], segmentation [71], object counting [31,66] and viewpoint estimation [74]. Our work is mostly related to StarMap [74], which proposes category-agnostic 3D keypoint representations encoded with canonical view locations. However, StarMap is only applicable for rigid objects (*e.g.* furniture), and relies on several expensive 3D CAD models of the target category to identify the predicted keypoint proposals. In comparison, CAPE aims at predicting 2D poses of any object category (both rigid and flexible) according to any manual keypoint definition given by one or a few support images.

## 2.3 Few-shot Learning

Few-shot learning [32] aims at learning novel classes using only a few examples. Recent few-shot learning approaches can be roughly classified into three categories, *i.e.* metric-learning-based approaches [48,54,64], meta-learning-based approaches [11,43], and data-augmentation-based approaches [14]. *Metric-learning-based Approaches.* Prototypical networks [48] learn the prototype (embedding features) of each class in the support data and then classify query data as the class whose prototype is the “nearest”. *Meta-learning-based Approaches.* Model-agnostic meta-learning [11] and LSTM-based meta-learner [43] aim at searching for a set of good initialization weights, such that the classifier can rapidly generalize to novel tasks by fine-tuning on only a few support samples. *Data-augmentation-based Approaches.* [14,56] generate synthetic examples of novel classes to improve the performance by using these synthetic examples for re-training. Our approach belongs to metric-learning-based approaches. It is the first framework towards CAPE. Besides, Keypoint Interaction Module (KIM) is specifically designed for CAPE to capture both the relationship among different keypoints and the relationship between support and query images.

# 3 Class-Agnostic Pose Estimation (CAPE)

## 3.1 Problem Definition

This paper introduces a novel task, termed class-agnostic pose estimation (CAPE). Unlike existing pose estimation tasks that predict keypoints of a single *known/seen* (super-)category, CAPE requires a single model to detect keypoints of arbitrary category. More specifically, given one or a few support samples with keypoint definition of an *unseen* category, object keypoints of this category can be detected without labeling large-scale supervisions and retraining models, significantly reducing the cost of data annotation and parameter tuning.

In order to validate the generalization capacity of CAPE models on unseen categories, they are trained on the *base* categories but evaluated on *novel* categories. The base categories and the novel categories are mutually exclusive, where the novel categories on the test set do not appear in the training data. During

testing, CAPE models are provided with  $K$  labeled support samples of an unseen category. The models are required to detect the poses of the query samples that are of the same category as the support samples. In this sense, CAPE task can be viewed as a  $K$ -shot pose estimation problem. Especially, when  $K = 1$ , it is one-shot pose estimation.

### 3.2 POse Matching Network (POMNet)

Traditional pose estimators can be applied to neither the unseen object categories nor different keypoint definitions of the same class (*e.g.* 19-keypoint human face definition and 68-keypoint human face definition). To achieve CAPE, we formulate the task as a matching problem and propose a novel framework termed POse Matching Network (POMNet). POMNet works by computing the matching similarity between the reference support keypoint features and the query image features at each location. Therefore, POMNet is capable of handling various categories with different keypoint numbers and definitions. As shown in Fig. 3, POMNet consists of three parts, *i.e.* the feature extractors ( $\Theta_S$  and  $\Theta_Q$ ), Keypoint Interaction Module (KIM), and Matching Head (MH).

**Feature Extractor.** We employ two parallel feature extractors to extract the support keypoint features and the query image features. In our implementation, ResNet-50 [15] pre-trained on ImageNet dataset is used as the backbone.

For the support image  $I_S$ , the feature extractor  $\Theta_S$  is utilized to extract the support image features  $\mathcal{F}_S = \Theta_S(I_S)$ . The keypoint annotations of the support sample are provided in the heatmap representations. We denote the ground-truth heatmaps of the support sample as  $H_S^*$ , and  $H_S^{*j} \in \mathbb{R}^{H \times W \times 1}$  represents the heatmap of the  $j_{th}$  keypoint. Given the support image features and the ground-truth heatmaps of the support sample, we can obtain the corresponding keypoint features as follows.

$$\hat{\mathcal{F}}_S^j = AvgPool(Upsample(\mathcal{F}_S) \otimes H_S^{*j}), \quad j = 1, 2, \dots, J \quad (1)$$

where  $\mathcal{F}_S \in \mathbb{R}^{h \times w \times c}$  and  $\hat{\mathcal{F}}_S^j \in \mathbb{R}^{1 \times 1 \times c}$  denote the support image features and the  $j_{th}$  keypoint features respectively.  $Upsample()$  is the up-sampling operation that reshapes the support image features to the same size of the corresponding heatmaps.  $\otimes$  denotes pixel-wise multiplication.  $AvgPool()$  is the average pooling operation that aggregates the support image features around the ground-truth keypoint position via weighted mean.  $J$  is the number of reference keypoints.

For the query image  $I_Q$ , we follow a similar pipeline and apply the feature extractor  $\Theta_Q$  to extract the query image features  $\mathcal{F}_Q = \Theta_Q(I_Q)$ . We collapse the spatial dimensions of the query image features and reshape them into a sequence. The extracted image features are then used to refine the support keypoint features in Keypoint Interaction Module (KIM) and to predict the keypoint localization in Matching Head (MH).

**Keypoint Interaction Module (KIM).** KIM targets at enhancing the support keypoint features through efficient attention mechanisms. We first reduce the channel dimension of support keypoint features by a fully-connected

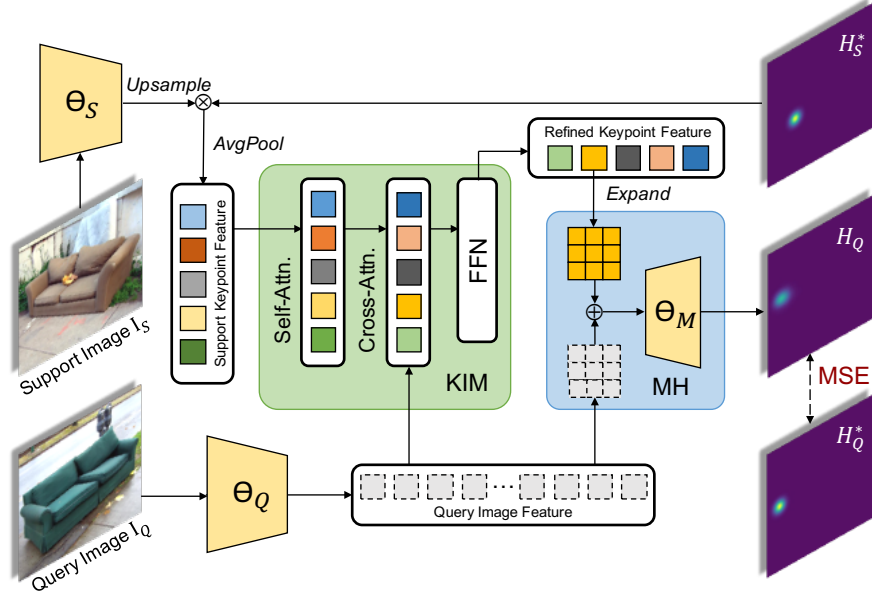


Fig. 3: Overview of POse Matching Network (POMNet). Feature extractors  $\Theta_S$  and  $\Theta_Q$  extract the support keypoint features and the query image features respectively. Keypoint Interaction Module (KIM) refines the keypoint features by message passing among keypoints and capturing the relationship between the query and support images. Matching Head (MH) integrates the refined keypoint features and the query image features to predict the keypoint localization in the query image. MSE loss is applied to supervise the model.

layer and input the features of different keypoints as a sequence. As the keypoint numbers of different categories are different, several dummy features with padding mask are added at the end to keep a fixed number  $L$  of input features ( $L = 100$  in our implementation), which enables KIM to adapt to various keypoint numbers. KIM has three transformer blocks, each of which consists of two major components, *i.e.* Self-Attn. and Cross-Attn. *Self-Attn.* The self-attention layer [53] learns to exchange information among keypoints and utilize inherent object structures. It allows the keypoint features to interact with each other, and aggregate these interactions using the attention weights. *Cross-Attn.* The keypoint features also interact with the query image features to align the feature representations and mitigate the representation gap. Specifically, a cross-attention layer [5] is applied to aggregate useful information in the query image. The keypoint features are input as query, and the flattened query image features are input as the key and the value. The channel dimension of the query image features are reduced to match the channel dimension of the keypoint features, and the sinusoidal position embedding [41, 53] is supplemented to the query image

features. A feed forward network (FFN) is also included following the common practice [53]. As a result, the support keypoint features are processed and refined by KIM,  $\{\hat{\mathcal{F}}_S^j\}_{j=1}^L = \text{KIM}(\{\mathcal{F}_S^j\}_{j=1}^L, \mathcal{F}_Q)$ . We exclude the dummy padding ones and obtain the refined keypoint features  $\{\bar{\mathcal{F}}_S^j\}_{j=1}^J$  by selecting the first  $J$  valid keypoint features, where  $J \leq L$ .

**Matching Head (MH).** Given the refined keypoint features as the reference, Matching Head (MH) targets at seeking the best matching positions in the query image that are encoded with heatmaps.

We expand the refined keypoint features to the same spatial shape as the query image features  $\mathcal{F}_Q$ . The expanded features are then concatenated with the query image features. Finally, a decoder  $\Theta_M$  is employed to estimate the keypoint heatmaps. This procedure can be formulated as follows.

$$H_Q^j = \Theta_M(\text{Expand}(\bar{\mathcal{F}}_S^j) \oplus \mathcal{F}_Q), \quad j = 1, 2, \dots, J. \quad (2)$$

where  $\oplus$  refers to the channel-wise concatenation.  $\text{Expand}()$  denotes the spatial expansion operation, *i.e.* copying the refined keypoint features spatially to fit in the spatial size of the query image features.  $H_Q^j$  is the predicted heatmap of the  $j$ th keypoint. The decoder  $\Theta_M$  consists of one  $3 \times 3$  convolutional layer, followed by deconvolutional layers for higher resolution as the common practice [62]. Pixel-wise mean squared error (MSE) loss is applied to supervise POMNet.

$$\mathcal{L}_{MSE} = \frac{1}{JHW} \sum_{j=1}^J \sum_p \|H_Q^j(p) - H_Q^{*j}(p)\|_2^2, \quad (3)$$

where  $H$  and  $W$  refer to the height and width of heatmaps.  $H_Q^j(p)$  and  $H_Q^{*j}(p)$  are the predicted and the ground-truth pixel intensity at the position  $p$ .

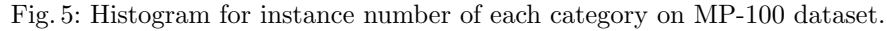
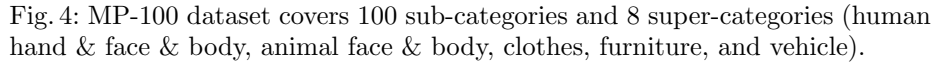
**Extension to K-shot.** When  $K$  ( $K > 1$ ) support images are available, we first extract the support keypoint features for each sample individually, and then calculate the mean among the  $K$  samples. The subsequent pipeline (including KIM and MH) is exactly the same as that of the 1-shot setting. With more support images, POMNet is able to capture more robust keypoint features to handle the intra-category variance and the ambiguity of the keypoint definition.

## 4 Multit-category Pose (MP-100) Dataset

Previous pose estimation datasets only consist of objects of one (super-)category and there are no existing datasets for CAPE task. We therefore construct the first large-scale pose dataset containing objects of multiple super-categories, termed Multi-category Pose (MP-100). In total, MP-100 dataset covers 100 sub-categories and 8 super-categories (human hand, human face, human body, animal body, animal face, clothes, furniture, and vehicle) as shown in Fig. 4.

Over 18K images and 20K annotations are collected from several popular 2D pose datasets, including COCO [30], 300W [45], AFLW [24], OneHand10K [55],





We split the collected 100 categories into train/val/test sets (70 for train, 10 for val, and 20 for test). Following the common settings, we form five splits whose test sets are non-overlapping and evaluate the average model performance on the five splits. In this case, each category is tested as novel one on different splits and the category bias is avoided. Moreover, to treat all categories equally, we try our best to balance the number of instances among different categories.

For the test set on each split, 2K samples are selected with 100 instances in each category. And for train/val set, 14K/2K samples are chosen for 70/10 categories respectively. As the number of instances available for different categories are extremely diverse and there are rare categories with less than 200 instances,

we minimize the standard deviation of the instance number of all the categories. During sample selection, for each category, we give preference to the instances with more valid keypoints labeled and larger image resolution. In Fig. 5, we demonstrate the histogram plot for the number of instances of each category on MP-100 dataset. The number of instances for each category is roughly balanced.

## 5 Experiments

### 5.1 Implementation Details

For each split on MP-100 dataset, we train POMNet on the train set, validate the performance on the val set, and finally evaluate the model on the test set. Note that the categories of the train/val/test set are mutually exclusive. During training, the support images and the query images of the same category are randomly paired. Each object of interest is cropped according to the bounding box and is resized to  $256 \times 256$ . Data augmentation with random scaling ( $[-15\%, 15\%]$ ) and random rotation ( $[-15^\circ, 15^\circ]$ ) is applied to improve the model generalization ability. The training is conducted on 8 GPUs with a batch size of 16 in each GPU for 210 epochs. We follow MMPose [9] to adopt Adam optimizer [23] with the base learning rate of  $1e-3$  and decay the learning rate to  $1e-4$  and  $1e-5$  respectively at the 170th and 200th epochs. During testing, we sample 3,000 random episodes for each novel category. Since there are 20 test categories for each split, we construct a total of 60,000 episodes for evaluation.

PCK (Probability of Correct Keypoint) is a popular metric for pose estimation. If the normalized distance between the predicted keypoint and the ground-truth keypoint is less than a certain threshold ( $\sigma$ ), it is considered correct.

$$\text{PCK} = \frac{1}{N} \sum_{i=1}^N \mathbf{1} \left( \frac{\|p_i - p_i^*\|_2}{d} \leq \sigma \right), \quad (4)$$

where  $p_i$  and  $p_i^*$  are predicted and ground-truth keypoint locations respectively.  $\mathbf{1}(\cdot)$  is the indicator function.  $d$  is the longest side of the ground-truth bounding box, which is used as the normalization term. The correct ratio of the overall  $N$  keypoints is calculated. In the experiments, we report the average PCK@0.2 ( $\sigma = 0.2$ ) of all the categories in each split. In order to minimize the category bias, the mean PCK result of all the 5 splits is also reported.

### 5.2 Benchmark Results on MP-100 Dataset

Class-Agnostic Pose Estimation (CAPE) is a new task that has not been tackled before. We tailor the existing few-shot learning baseline approaches, including Prototypical Networks [48], Finetune [38], and MAML [11] to address this new task. For fair comparisons, all baselines employ the same backbone network architecture (ResNet-50 [15]) as ours.

**Prototypical Networks (ProtoNet)** [48]. ProtoNet is a popular few-shot image classification approach, which constructs a prototype for each class and

Table 1: Comparisons with the baseline methods on MP-100 dataset under both 5-shot and 1-shot settings. POMNet significantly outperforms other approaches.

5-Shot	Split1	Split2	Split3	Split4	Split5	Mean (PCK)
ProtoNet [48]	60.31	53.51	61.92	58.44	58.61	58.56
MAML [11]	70.03	55.98	63.21	64.79	58.47	62.50
Fine-tune [38]	71.67	57.84	66.76	66.53	60.24	64.61
POMNet (Ours)	<b>84.72</b>	<b>79.61</b>	<b>78.00</b>	<b>80.38</b>	<b>80.85</b>	<b>80.71</b>
1-Shot	Split1	Split2	Split3	Split4	Split5	Mean (PCK)
ProtoNet [48]	46.05	40.84	49.13	43.34	44.54	44.78
MAML [11]	68.14	54.72	64.19	63.24	57.20	61.50
Fine-tune [38]	70.60	57.04	66.06	65.00	59.20	63.58
POMNet (Ours)	<b>84.23</b>	<b>78.25</b>	<b>78.17</b>	<b>78.68</b>	<b>79.17</b>	<b>79.70</b>

the query example is assigned to the class whose prototype is the “nearest”. To solve CAPE, we adapt ProtoNet to construct a prototype for each keypoint and find the location whose features are closest to the prototype in the query image. Unlike image classification, both the receptive fields and the spatial resolution of the features are critical for pose estimation. We empirically find that the features from Stage-3 achieve a good balance of these two factors among all the 4 stages.

**Fine-tune [38].** The model is first pre-trained using a combination of all base categories on the train set. During testing, the model is fine-tuned on the support images of the novel category before estimating the pose of the query images. To handle the problem of various number of keypoints, the model is designed to output the maximum number of keypoints among all the categories, *i.e.* 68 on MP-100 dataset, and only the few valid keypoints are supervised for each particular category during training and fine-tuning.

**Model-Agnostic Meta-Learning (MAML) [11].** Through meta training, the MAML model is explicitly trained to search for a good initialization such that its parameters can quickly adapt to the given category by fine-tuning on several support images. Similar to Fine-tune [38], the number of keypoints of the model is set as 68. During meta testing, the model can rapidly adapt to the novel categories given a few support images.

As shown in Table 1, our proposed POMNet shows superiority over the existing few-shot learning based approaches on the task of Class-Agnostic Pose Estimation (CAPE). We first conduct experimental comparisons under the 5-shot setting. We observe that ProtoNet [48] mostly relies on low-level appearance features and encounters difficulties in constructing a reliable prototype using only 5 samples for all the keypoints. It processes each type of keypoint individually and does not utilize the structural information, restricting its upper bound performance. MAML [11] and Fine-tune [38] adapt to the novel object category by fine-tuning on a few samples during testing. However, the limited number of samples makes it hard for the model to achieve good performance on the novel categories due to severe over-fitting or under-fitting problems. Our pro-

Table 2: Cross super-category evaluation (PCK). POMNet outperforms other methods. But there is still large room for improvement on the rare categories.

Method	Human Body	Human Face	Vehicle	Furniture
ProtoNet [48]	37.61	57.80	28.35	42.64
MAML [11]	51.93	25.72	17.68	20.09
Fine-tune [38]	52.11	25.53	17.46	20.76
POMNet (Ours)	<b>73.82</b>	<b>79.63</b>	<b>34.92</b>	<b>47.27</b>

posed POMNet considers the CAPE task as a matching problem, decoupling the model from the object category and the number of keypoints. In the meanwhile, KIM explicitly captures the relationship among keypoints and the structure of the object of interest. As a result, POMNet achieves 80.71 PCK on the novel categories under 5-shot settings, and outperforms the baseline methods by a large margin (over 25% improvement).

When the number of sample images decreases to one, the degeneration of our POMNet is only 1.3% (79.70 vs 80.71 PCK). This is presumably because POMNet captures the relationship among semantic keypoints and is more robust to occlusion and visual ambiguity. In comparison, ProtoNet requires building the prototype based on a single keypoint, thus is more sensitive to the appearance variation, resulting in a larger performance drop (44.78 vs 58.46 PCK).

### 5.3 Cross Super-Category Pose Estimation

In order to further evaluate the generalization ability, we conduct the cross super-category pose estimation evaluation with the “Leave-One-Out” strategy. That is, we train the model on all but one super-categories on MP-100 dataset, and evaluate the performance on the remaining one super-category. The super-categories to be evaluated include human body, human face, vehicle, and furniture.

As shown in Table 2, our proposed POMNet outperforms the baseline methods on all the super-categories, demonstrating stronger generalization ability. However, super-category generalization is challenging and there is still a large room for improvement. We notice that all the methods perform poorly on the super-categories of vehicle and furniture. This is possibly because these categories are very different from the training ones and the extracted features are not discriminative enough. There are a great number of invisible keypoints for vehicle, and the intra-class variation between images is large for furniture, making these two super-categories more challenging. Solving CAPE requires to handle occlusion and intra-class appearance variation, and extract more discriminative features for unseen categories. We will explore these directions in the future.

### 5.4 Ablation Study

**Effect of model components.** Table 3 shows the effect of Keypoint Interaction Module (KIM) and Matching Head (MH). Comparing #1 and #5, we find

Table 3: Ablation study of proposed components on MP-100 Split1 under 1-shot setting. KIM and MH significantly improve the model performance.

	Self-Atten.	Cross-Atten.	Matching Head	PCK
#1			✓	74.40
#2	✓	✓		79.19
#3		✓	✓	80.76
#4	✓		✓	82.92
#5	✓	✓	✓	84.23

that KIM significantly improves the CAPE model performance (13.2% improvement). #3 and #4 show the effect of self-attention and cross-attention design, respectively. Especially, the 11.5% gain from #1 to #4 shows that message passing among keypoints by self-attention greatly benefits keypoint localization. Comparison between #2 and #5 verifies the necessity of MH. #2 replaces MH by matrix multiplication between support keypoint features and query image features. It collapses the channel dimension to 1 for each keypoint, causing undesirable information loss required for precise localization.

Table 4: **Left:** Effect of training category number (“#Train”) under 1-shot setting. Evaluation is conducted on a novel category (“human body”). **Right:** Both training and testing are on “human body” only.

#Train	1	9	49	99	Oracle	SBL-Res50 [60]
PCK	39.32	55.74	70.46	73.82	89.79	89.76

**Effect of training category number.** As shown in Table 4 **Left**, more training categories leads to better generalizability to the novel category, which validates the necessity of MP-100 dataset and the rationality of our experiments.

**Sanity check.** We perform traditional one class pose estimation as a sanity check. In Table 4 **Right**, “Oracle” means POMNet trained and tested on the same category (“human body”) only. It performs comparable with SBL-Res50 [15], which demonstrates the correctness of our design choices.

## 5.5 Qualitative Results

In Fig. 6, we qualitatively evaluate the generalization ability of POMNet to the novel categories on MP-100 test sets. Our method is robust to perspective variation and appearance diversity. Typical failure cases include appearance ambiguity (the first two examples) and severe occlusion (the 3rd example).

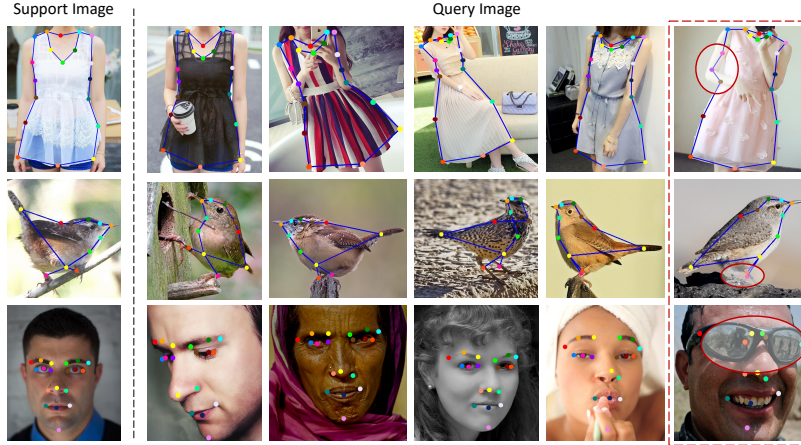


Fig. 6: Qualitative results of POMNet on unseen categories. The first column shows the manually annotated support samples, and the others are the predicted query samples. The last column shows some failure cases (in RED circles).

## 6 Conclusions and Limitations

This paper introduces a novel task, termed Category-Agnostic Pose Estimation (CAPE). The idea of CAPE can benefit a wide range of application scenarios. It would not only promote the development of pose estimation (*e.g.* pseudo-labeling for novel categories), but also enable the researchers in the other fields to detect keypoints of objects they are interested in (*e.g.* plants). Besides, it may also make broader positive impacts for other computer vision tasks. For example, CAPE models can be developed for keypoint-based object tracking, contour-based instance segmentation, and graph matching. To achieve this goal, we propose the first CAPE framework, POse Matching Network (POMNet), and the first dataset for CAPE task, Multi-category Pose (MP-100). Experiments show that POMNet significantly outperforms the other approaches on MP-100 dataset. However, there are still many remaining challenges, *e.g.* the generalization performance on rare categories, intra-class appearance variation, self-occlusion, and appearance ambiguity. In conclusion, CAPE, as an important yet challenging task, is worth more research attention and further exploration.

**Acknowledgement.** This work is supported in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants (Nos. 14202217, 14203118, 14208619), in part by Research Impact Fund Grant No. R5001-18. Ping Luo is supported by the General Research Fund of HK No.27208720, No.17212120, and No.17200622. Wanli Ouyang is supported by the Australian Research Council Grant DP200103223, Australian Medical Research Future Fund MRFAI000085, CRC-P Smart Material Recovery Facility (SMRF) – Curby Soft Plastics, and CRC-P ARIA - Bionic Visual-Spatial Prosthesis for the Blind.



## References

1. Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., Schiele, B.: Posetrack: A benchmark for human pose estimation and tracking. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018) [4](#)
2. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: IEEE Conf. Comput. Vis. Pattern Recog. (2014) [1](#), [4](#)
3. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? In: Int. Conf. Comput. Vis. (2017) [4](#)
4. Cao, J., Tang, H., Fang, H.S., Shen, X., Lu, C., Tai, Y.W.: Cross-domain adaptation for animal pose estimation. In: Int. Conf. Comput. Vis. (2019) [4](#)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Eur. Conf. Comput. Vis. (2020) [7](#)
6. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018) [4](#)
7. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020) [4](#)
8. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2017) [4](#)
9. Contributors, M.: Openmmlab pose estimation toolbox and benchmark. <https://github.com/open-mmlab/mmpose> (2020) [10](#)
10. Duan, H., Lin, K.Y., Jin, S., Liu, W., Qian, C., Ouyang, W.: Trb: a novel triplet representation for understanding 2d human body. In: Int. Conf. Comput. Vis. (2019) [4](#)
11. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML (2017) [5](#), [10](#), [11](#), [12](#)
12. Ge, Y., Zhang, R., Wang, X., Tang, X., Luo, P.: Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In: IEEE Conf. Comput. Vis. Pattern Recog. (2019) [4](#), [9](#)
13. Graving, J.M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B.R., Couzin, I.D.: Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife* (2019) [9](#)
14. Hariharan, B., Girshick, R.: Low-shot visual recognition by shrinking and hallucinating features. In: Int. Conf. Comput. Vis. (2017) [5](#)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. (2016) [6](#), [10](#), [13](#)
16. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: Int. Conf. Comput. Vis. (2013) [4](#)
17. Jiang, S., Liang, S., Chen, C., Zhu, Y., Li, X.: Class agnostic image common object detection. *IEEE Trans. Image Process.* **28**(6), 2836–2846 (2019) [5](#)
18. Jin, S., Liu, W., Ouyang, W., Qian, C.: Multi-person articulated tracking with spatial and temporal embeddings. In: IEEE Conf. Comput. Vis. Pattern Recog. (2019) [4](#)
19. Jin, S., Liu, W., Xie, E., Wang, W., Qian, C., Ouyang, W., Luo, P.: Differentiable hierarchical graph grouping for multi-person pose estimation. In: Eur. Conf. Comput. Vis. (2020) [4](#)

20. Jin, S., Ma, X., Han, Z., Wu, Y., Yang, W., Liu, W., Qian, C., Ouyang, W.: Towards multi-person pose tracking: Bottom-up and top-down methods. In: Int. Conf. Comput. Vis. Worksh. (2017) [4](#)
21. Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., Luo, P.: Whole-body human pose estimation in the wild. In: Eur. Conf. Comput. Vis. (2020) [4](#)
22. Khan, M.H., McDonagh, J., Khan, S., Shahabuddin, M., Arora, A., Khan, F.S., Shao, L., Tzimiropoulos, G.: Animalweb: A large-scale hierarchical dataset of annotated animal faces. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020) [4](#), [9](#)
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Int. Conf. Learn. Represent. (2015) [10](#)
24. Kostinger, M., Wohlhart, P., Roth, P., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: Int. Conf. Comput. Vis. Worksh. (2011) [4](#), [8](#)
25. Labuguen, R., Matsumoto, J., Negrete, S.B., Nishimaru, H., Nishijo, H., Takada, M., Go, Y., Inoue, K.i., Shibata, T.: Macaquepose: A novel “in the wild” macaque monkey pose dataset for markerless motion capture. *Frontiers in behavioral neuroscience* (2021) [4](#), [9](#)
26. Li, J., Bian, S., Zeng, A., Wang, C., Pang, B., Liu, W., Lu, C.: Human pose regression with residual log-likelihood estimation. In: Int. Conf. Comput. Vis. (2021) [4](#)
27. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: IEEE Conf. Comput. Vis. Pattern Recog. (2019) [4](#)
28. Li, S., Li, J., Tang, H., Qian, R., Lin, W.: Atrw: A benchmark for amur tiger re-identification in the wild. In: ACM Int. Conf. Multimedia (2020) [4](#)
29. Li, Y., Zhang, S., Wang, Z., Yang, S., Yang, W., Xia, S.T., Zhou, E.: Tokenpose: Learning keypoint tokens for human pose estimation. *arXiv preprint arXiv:2104.03516* (2021) [4](#)
30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Eur. Conf. Comput. Vis. (2014) [4](#), [8](#)
31. Lu, E., Xie, W., Zisserman, A.: Class-agnostic counting. In: ACCV (2018) [5](#)
32. Lu, J., Gong, P., Ye, J., Zhang, C.: Learning from very few samples: A survey. *arXiv preprint arXiv:2009.02653* (2020) [5](#)
33. Mao, W., Ge, Y., Shen, C., Tian, Z., Wang, X., Wang, Z.: Tfpote: Direct human pose estimation with transformers. *arXiv preprint arXiv:2103.15320* (2021) [4](#)
34. Mathis, A., Biasi, T., Schneider, S., Yuksekgonul, M., Rogers, B., Bethge, M., Mathis, M.W.: Pretraining boosts out-of-domain robustness for pose estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2021) [4](#)
35. Moon, G., Yu, S.I., Wen, H., Shiratori, T., Lee, K.M.: Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In: Eur. Conf. Comput. Vis. (2020) [4](#)
36. Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., Theobalt, C.: Generated hands for real-time 3d hand tracking from monocular rgb. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018) [4](#)
37. Mueller, F., Mehta, D., Sotnychenko, O., Sridhar, S., Casas, D., Theobalt, C.: Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In: Int. Conf. Comput. Vis. (2017) [4](#)
38. Nakamura, A., Harada, T.: Revisiting fine-tuning for few-shot learning. *arXiv preprint arXiv:1910.00216* (2019) [10](#), [11](#), [12](#)



39. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Eur. Conf. Comput. Vis. (2016) [4](#)
40. Nie, X., Feng, J., Zhang, J., Yan, S.: Single-stage multi-person pose machines. In: Int. Conf. Comput. Vis. (2019) [4](#)
41. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D.: Image transformer. In: ICML (2018) [7](#)
42. Pereira, T.D., Aldarondo, D.E., Willmore, L., Kislin, M., Wang, S.S.H., Murthy, M., Shaevitz, J.W.: Fast animal pose estimation using deep neural networks. Nature methods (2019) [9](#)
43. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: Int. Conf. Learn. Represent. (2017) [5](#)
44. Reddy, N.D., Vo, M., Narasimhan, S.G.: Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicles. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018) [1](#), [4](#), [9](#)
45. Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: database and results. Image and Vision Computing (2016) [4](#), [8](#)
46. Shen, J., Zafeiriou, S., Chrysos, G.G., Kossaifi, J., Tzimiropoulos, G., Pantic, M.: The first facial landmark tracking in-the-wild challenge: Benchmark and results. In: Int. Conf. Comput. Vis. Worksh. (2015) [4](#)
47. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: IEEE Conf. Comput. Vis. Pattern Recog. (2017) [4](#)
48. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Adv. Neural Inform. Process. Syst. (2017) [5](#), [10](#), [11](#), [12](#)
49. Song, X., Wang, P., Zhou, D., Zhu, R., Guan, C., Dai, Y., Su, H., Li, H., Yang, R.: Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving. In: IEEE Conf. Comput. Vis. Pattern Recog. (2019) [4](#)
50. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2019) [2](#), [4](#)
51. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: Int. Conf. Comput. Vis. (2017) [4](#)
52. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: IEEE Conf. Comput. Vis. Pattern Recog. (2014) [4](#)
53. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Adv. Neural Inform. Process. Syst. (2017) [7](#), [8](#)
54. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: Adv. Neural Inform. Process. Syst. (2016) [5](#)
55. Wang, Y., Peng, C., Liu, Y.: Mask-pose cascaded cnn for 2d hand pose estimation from single color image. IEEE Transactions on Circuits and Systems for Video Technology (2018) [4](#), [8](#)
56. Wang, Y.X., Girshick, R., Hebert, M., Hariharan, B.: Low-shot learning from imaginary data. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018) [5](#)
57. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: IEEE Conf. Comput. Vis. Pattern Recog. (2016) [4](#)
58. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-UCSD Birds 200. Tech. Rep. CNS-TR-2010-001, California Institute of Technology (2010) [4](#), [9](#)

59. Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., Wang, W., Zhou, S., Lin, G., Fu, Y., et al.: Ai challenger: A large-scale dataset for going deeper in image understanding. arXiv preprint arXiv:1711.06475 (2017) [4](#)
60. Wu, J., Xue, T., Lim, J.J., Tian, Y., Tenenbaum, J.B., Torralba, A., Freeman, W.T.: Single image 3d interpreter network. In: Eur. Conf. Comput. Vis. (2016) [4](#), [9](#)
61. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: A boundary-aware face alignment algorithm. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018) [4](#)
62. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Eur. Conf. Comput. Vis. (2018) [4](#), [8](#)
63. Xu, L., Guan, Y., Jin, S., Liu, W., Qian, C., Luo, P., Ouyang, W., Wang, X.: Vipnas: Efficient video pose estimation via neural architecture search. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021) [4](#)
64. Yang, F.S.Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018) [5](#)
65. Yang, S., Quan, Z., Nie, M., Yang, W.: Transpose: Towards explainable human pose estimation by transformer. arXiv preprint arXiv:2012.14214 (2020) [4](#)
66. Yang, S.D., Su, H.T., Hsu, W.H., Chen, W.C.: Class-agnostic few-shot object counting. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2021) [5](#)
67. Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., Tao, D.: Ap-10k: A benchmark for animal pose estimation in the wild. arXiv preprint arXiv:2108.12617 (2021) [1](#), [4](#), [9](#)
68. Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., Wang, J.: Hrformer: High-resolution transformer for dense prediction. arXiv preprint arXiv:2110.09408 (2021) [4](#)
69. Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., Shen, J.: The menpo facial landmark localisation challenge: A step towards the solution. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. (2017) [4](#)
70. Zeng, W., Jin, S., Liu, W., Qian, C., Luo, P., Ouyang, W., Wang, X.: Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022) [4](#)
71. Zhang, C., Lin, G., Liu, F., Yao, R., Shen, C.: Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5217–5226 (2019) [5](#)
72. Zhang, S.H., Li, R., Dong, X., Rosin, P., Cai, Z., Han, X., Yang, D., Huang, H., Hu, S.M.: Pose2seg: Detection free human instance segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2019) [4](#)
73. Zhao, J., Li, J., Cheng, Y., Sim, T., Yan, S., Feng, J.: Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In: ACM Int. Conf. Multimedia (2018) [4](#)
74. Zhou, X., Karpur, A., Luo, L., Huang, Q.: Starmap for category-agnostic keypoint and viewpoint estimation. In: Eur. Conf. Comput. Vis. pp. 318–334 (2018) [5](#)
75. Zimmermann, C., Brox, T.: Learning to estimate 3d hand pose from single rgb images. In: Int. Conf. Comput. Vis. (2017) [4](#)
76. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In: Int. Conf. Comput. Vis. (2019) [4](#)